National Library
of Canada

Bibliothèque nationale
du Canada

Canadian Theses Service

Service des thèses canadiennes

Ottawa, Canada
K1A 0N4

NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30, and subsequent amendments.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30, et ses amendements subséquents.

NL-339 (r. 88/04) c

Canada

National Library of Canada
Bibliothèque nationale du Canada

Canadian Theses Service
Service des thèses canadiennes

Ottawa, Canada
K1A 0N4

The author has granted an irrevocable non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of his/her thesis by any means and in any form or format, making this thesis available to interested persons.

The author retains ownership of the copyright in his/her thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without his/her permission.

L'auteur a accordé une licence irrévocable et non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de sa thèse de quelque manière et sous quelque forme que ce soit pour mettre des exemplaires de cette thèse à la disposition des personnes intéressées.

L'auteur conserve la propriété du droit d'auteur qui protège sa thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

ISBN 0-315-55489-4

Canada

The University of Alberta

MONTE CARLO COMPARISON OF ANOVA AND LOGLINEAR ANALYSIS USING BERNOULLI DATA

by

William Petroske ©

A thesis
submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree
of Master of Science

Department of Computing Science

Edmonton, Alberta
Fall, 1989

THE UNIVERSITY OF ALBERTA

RELEASE FORM

NAME OF AUTHOR: William Petroske

TITLE OF THESIS: Monte Carlo Comparison of ANOVA and Loglinear Analysis Using Bernoulli Data

DEGREE FOR WHICH THIS THESIS WAS PRESENTED: Master of Science

YEAR THIS DEGREE GRANTED: 1989

Permission is hereby granted to The University of Alberta Library to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly or scientific research purposes only.

The author reserves other publication rights, and neither the thesis nor extensive extracts from it may be printed or otherwise reproduced without the author's written permission.

(Signed) William J. Petroske

Permanent Address:
4134 SW Washouga Ave.
Portland, Oregon
USA 97201

Dated 29 September 1989

THE UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES AND RESEARCH

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research, for acceptance, a thesis entitled **Monte Carlo Comparison of ANOVA and Loglinear Analysis Using Bernoulli Data** submitted by **William Petroske** in partial fulfillment of the requirements for the degree of **Master of Science.**

Kellogg Wilson
Co-supervisor

Keith Smillie
Co-supervisor

Barry Gee

Jerome N. Sheahan

Date September 29, 1989

ABSTRACT

Logit models provide a means of modelling dependency among variables. For the completely categorical case, loglinear models may be used to fit corresponding logit models. Usually the response variable is dichotomous so that the logit formulation can be in terms of simple odds of the two categories. The same data analyzed by logit models where the response variable is dichotomous may be viewed as outcomes in a series of Bernoulli trials and arranged as responses in a crossed ANOVA layout. In the case of Bernoulli type data, it has been observed that fixed-effects, balanced ANOVA is robust when the treatment group probabilities of success are not extreme and the sample size is not too small.

This study focussed on empirical comparisons of loglinear analysis (logit) and ANOVA. An assortment of 2x2 and 2x2x2 tables (ANOVA layout perspective) that vary in the group treatment probabilities of success were examined via Monte Carlo simulations. The simulations primarily investigated completely null models and models with main effects present for both types of modelling (tests for first order interactions were included). The study was based mostly upon the Type I error levels and power results of the F-test for ANOVA and the conditional likelihood ratio test G^2 for loglinear fits. Small and moderate sample sizes were used along with a full range of probabilities. In addition, results were obtained to examine the effects of using the Freeman-Tukey arcsine transformation in ANOVA and the effects of adding 0.5 to all elementary cell totals for loglinear analysis. The overall goodness-of-fit statistics for loglinear modelling, the Pearson X^2 and the likelihood ratio G^2, were compared. Also, stepwise forward and stepwise backward model selection strategies were contrasted.

Acknowledgements

Thanks to the gang. Thanks to Edith Drummond. Thanks to those who prefer maybe not to be mentioned (like George, Don, and Dave).

Table of Contents

Chapter Page

Chapter 1: Introduction 1

Chapter 2: Analysis of Variance 3

2.1. The Basics of ANOVA 3

2.2. Fixed-Effects Two-Way and Three-way ANOVA 6

2.3. Assumptions Underlying ANOVA F-Test 8

2.4. Model Selection 10

Chapter 3: The Loglinear Model 12

3.1. History 12

3.2. The Model 13

3.3. Sampling Schemes 15

3.4. Estimation and Model Selection 16

3.5. Logit Model 19

3.6. ANOVA/Loglinear Analogy—Example 21

Chapter 4: Related Studies 23

4.1. Studies Related to Loglinear Analysis 23

4.2. Studies Related to ANOVA 25

4.3. Comparisons of Loglinear Analysis and ANOVA 27

Chapter 5: Simulation Design 31

5.1. Hypothesis Testing 31

5.2. Kolomogorov-Smirnov Test 32

5.3. Type I Error Rates 33

5.4. Type II Error Rates (Power) 34

5.5. Compared Orderings of Sample Tables 35

5.6. Simulation Design 37

5.6.1. Aim and Scope 37

5.6.2. Environment 39

5.6.3. Table Dimensionality 40

5.6.4. Cell Sample Size 42

5.6.5. Data Transformations 43

5.6.5.1. Arcsine Transformation 43

5.6.5.2. Adding 0.5 to All Elementary Cell Totals 44

5.7. Models and Probability Structures 45

5.7.1. Completely Null Model 45

5.7.2. Main Effects for One Variable 47

5.7.3. Full Main Effects Model 49

Chapter 6: Simulation Results 52

6.1. Interpretation of the Graphs 53

6.2. Tabled Results 55

6.3. Arcsine Transformation 57

6.3.1. Type I Error – Upper Percentage Points 57

6.3.2. Power 73

6.3.3. Further Discussion 75

6.4. Addition of 0.5 to Cell Totals 79

6.5. G^2 Versus X^2 85

6.6. Comparison of ANOVA and Loglinear Results 88

6.6.1. Type I Error – No Effects Present 89

6.6.2. Models With Effects Present – Power and Type I Error 99

6.6.2.1. Type I Results – Effects Present 100

6.6.2.2. Power 103

6.7. Other Measures 106

Chapter 7: Conclusion 107

References 109

A1: Graphs 114

A2: Additional Tables 150

List of Tables

Table Page

2.1 General layout for one-way ANOVA design. 4

2.2 Table layout and parametric equation for 2×2 ANOVA examined in this study (T1 and T2 denote treatments). 6

2.3 Table layout and parametric model for 2×2×2 ANOVA examined in this study (T1 and T2 denote treatments). 7

3.1 Example 2×2 table of probabilities. 13

3.2 Example of Bernoulli trial data for 2×2 ANOVA. 21

3.3 Example contingency table for data in Table 3.2 (above). 22

4.1 Magidson example 1: row effects (very small for both probability and odds), column effects (small for both probability and odds), and interaction effects (none in probabilities, very small in odds). 29

4.2 Magidson example 2: row effects (very small in probabilities, large in odds), column effects (large in probabilities, very large in odds), and interaction effects (none in probabilities, large in odds). 29

4.3 Magidson example 3: row effects (small in probabilities, large in odds), column effects (very large in probabilities, extremely large in odds), and interaction effects (large in probabilities, none in odds). 29

4.4 Magidson example 4: Every value of X in the additive probability model must show an interaction effect whereas the odds model shows no interaction effect for $X = 144$ (odds) or 0.993 (probability). 29

5.1 Null model cell probabilities for 2×2 tables. 46

5.2 Null model cell probabilities for $2 \times 2 \times 2$ tables. 46

5.3 Cell probabilities of 2×2 tables for row main effects. 47

5.4 Cell probabilities of $2 \times 2 \times 2$ tables for row main effects. 49

5.5 Cell probabilities for 2×2 tables with both row and column effects present. 50

5.6 Cell probabilities for $2 \times 2 \times 2$ tables with both row, column, and layer effects present. 51

6.1 2×2 table probabilities (A, B, C, D) and $2 \times 2 \times 2$ table probabilities (A, B, C, D, E, F, G, H) 55

6.2 Summary of probabilities used in test cases. Effects are noted in terms of additive probability model. Cases marked with * have large log-odds interaction effects. In all other cases the presence of log-odds effects corresponds with the labels in column two. 56

6.3 Standard deviations from nominal Type I error rates for ANOVA test of null hypothesis of no row effects 58

6.4 Empirical Type I error rates for two dimensional models at nominal 5 percent significance level for null hypotheses of no column and interaction effects while row effects are present 69

6.5 Empirical Type I error rates (nominal 5 percent level) for true null hypothesis of no interaction effects (additive probability model) for two dimensional cases with both row and column effects present 70

6.6 Empirical Type I error rates for three dimensional cases where a main effect is present. 71

6.7 Power comparison at nominal 0.05 significance level between ANOVA F for raw scores and ANOVA F after arcsine transformation of scores. 74

6.8 Power of ANOVA F and Conditional G^2 for three dimensional cases where all main effects are present. 75

6.9 Standard deviations of G^2 and X^2 from nominal Type I error rates after cell counts have been adjusted by addition of 0.5 to each cell (results without adjustment are in Table 6.14). 80

6.10 Power for two dimensional cases with row effects present; null hypothesis of no row effects. 81

6.11 Power for two dimensional cases with row and column effects present; null hypothesis of no effects(LL overall fit), no row effects(ANOVA, conditional G^2). 82

6.12 Power for two dimensional cases with row and column effects present; null hypothesis of no column effects. 83

6.13 Power for 3 dimensional cases with row,column, layer effects present (cases 6.1,6.2), and cases with only row and column effects (4.1,4.2,4.3): tests for row effects (ANOVA, conditional G^2). 84

6.14 Standard deviations from nominal Type I error rates for X^2 and G^2 as measures of overall goodness-of-fit for null hypothesis (true) of no effects. 87

6.15 Type I error results for case 1.1: two factors, n=10 per cell, success probability=0.05 per cell. 90

6.16 Type I error results for case 1.2: two factors, n=10 per cell, success probability=0.1 per cell. 90

6.17 Type I error results for case 1.3: two factors, n=10 per cell, success probability=0.175 per cell. 91

6.18 Type I error results for case 1.4: two factors, n=10 per cell, success probability=0.25 per cell. 91

6.19 Type I error results for case 1.5: two factors, n=10 per cell, success probability=0.375 per cell. 92

6.20 Type I error results for case 1.6: two factors, n=10 per cell, success probability=0.5 per cell. 92

6.21 Type I error results for case 1.1: two factors, n=40 per cell, success probability=0.05 per cell. 93

6.22 Type I error results for case 1.2: two factors, n=40 per cell, success probability=0.1 per cell. 93

6.23 Type I error results for case 1.3: two factors, n=40 per cell, success probability=0.175 per cell. 94

6.24 Type I error results for case 1.4: two factors, n=40 per cell, success probability=0.25 per cell. 94

6.25 Type I error results for case 1.5: two factors, n=40 per cell, success probability=0.375 per cell. 95

6.26 Type I error results for case 1.6: two factors, n=40 per cell, success probability=0.5 per cell. 95

6.27 Logit models fitted in Monte Carlo study. 101

6.28 Summary of models used in calculations of conditional G^2 forward and conditional G^2 backward. 101

6.29 Power comparison at nominal 0.05 significance level between ANOVA F for raw scores and conditional G^2 forward. 104

6.30 Power: ANOVA F and G^2 backward. 104

6.31 Power comparison at nominal 0.05 significance level between conditional G^2 statistics given by the forward and backward stepwise model selection strategies of loglinear analysis. 105

A2.1 Type I error levels for case 2.1: Three factors, success probability of 0.1 per cell, 10 and 40 observations per cell. 151

A2.2 Type I error levels for case 2.2: three factors, success probability of 0.25 per cell, 10 and 40 observations per cell. 151

A2.3 Type I error levels for case 2.3: three factors, success probability of 0.5 per cell, 10 and 40 observations per cell. 151

List of Figures

Figure Page

6.1 Example plot of nominal versus empirical cumulative 53

6.2 ANOVA F with arcsine transformation; ANOVA F without arcsine transformation; conditional G^2 backwards: testing for row effects, 40 observations per cell, 2×2×2 table, $p = 0.5$ (case 2.3). 61

6.3 Conditional G^2 backward and ANOVA F with arcsine transformation testing for row effects: 2×2×2 table, 40 observations per cell, $p = 0.1$ (case 2.1). 62

6.4 ANOVA F testing for column effects with no effects present: $p = 0.1$, 10 observations per cell, 2×2 table (case 1.2). 63

6.5 ANOVA F with arcsine transformation testing for column effects with no effects present: $p = 0.1$, 10 observations per cell, 2×2 table (case 1.2). 64

6.6 ANOVA F testing for column effects with no effects present: $p = 0.25$, 10 observations per cell, 2×2 table (case 1.4). 65

6.7 ANOVA F with arcsine transformation testing for column effects with no effects present: $p = 0.25$, 10 observations per cell, 2×2 table (case 1.4). 66

6.8 ANOVA F testing for column effects with no effects present: $p = 0.5$, 10 observations per cell, 2×2 table (case 1.6). 67

6.9 ANOVA F testing with arcsine transformation for column effects with no effects present: $p = 0.5$, 10 observations per cell, 2×2 table (case 1.6). 68

6.10 Graph of ratio of actual to asymptotic variance, $\sigma_\theta^2/\sigma_\infty^2$, versus binomial probability, p, for $n = 10$ (adapted from Mosteller and Youtz) 76

6.11 Graph of number of standard deviations from nominal 5% level versus binomial probability, p, for $n = 10$ 78

A1.1 Overall goodness-of-fit X^2 testing no effects present (true model): $p = 0.1$, 40 observations per cell, 2×2×2 table (case 2.1). 114

A1.2 Overall goodness-of-fit X^2 testing no effects present (true model): $p = 0.25$, 40 observations per cell, 2×2×2 table (case 2.2). 115

A1.3 Overall goodness-of-fit X^2 testing no effects present (true model): $p = 0.5$, 40 observations per cell, 2×2×2 table (case 2.3). 116

A1.4 Overall goodness-of-fit X^2 testing no effects present (true model): $p = 0.1$, 10 observations per cell, 2×2 table (case 1.2). 117

A1.5 Overall goodness-of-fit X^2 testing no effects present (true model): $p = 0.25$, 10 observations per cell, 2×2 table (case 1.4). 118

A1.6 Overall goodness-of-fit X^2 testing no effects present (true model): $p = 0.5$, 10 observations per cell, 2×2 table (case 1.6). 119

A1.7 ANOVA F testing for row effects with no effects present: $p = 0.1$, 40 observations per cell, 2×2×2 table (case 2.1). 120

A1.8 ANOVA F testing for row effects with no effects present: $p = 0.25$, 40 observations per cell, 2×2×2 table (case 2.2). 121

A1.9 ANOVA F testing for row effects with no effects present: $p = 0.5$, 40 observations per cell, 2×2×2 table (case 2.3). 122

A1.10 ANOVA F testing with arcsine transformation for row-column interaction effects with no effects present: $p = 0.5$, 40 observations per cell, 2×2×2 table (case 2.3). 123

A1.11 G^2 overall goodness-of-fit: 2×2×2 table, 10 observations per cell, $p = 0.1$, $p = 0.25$, $p = 0.5$ (cases 2.1, 2.2, 2.3, respectively). 124

A1.12 G^2 overall goodness-of-fit after addition of 0.5 to each cell total: 2×2×2 table, 10 observations per cell, $p = 0.1$, $p = 0.25$, $p = 0.5$ (cases 2.1, 2.2, 2.3, respectively). 125

A1.13 X^2 overall goodness-of-fit: 2×2×2 table, 10 observations per cell, $p = 0.1$, $p = 0.25$, $p = 0.5$ (cases 2.1, 2.2, 2.3, respectively). 126

A1.14 X^2 overall goodness-of-fit and ANOVA F for row effects: 2×2×2 table, 10 observations per cell, $p = 0.1$ (case 2.1). 127

A1.15 X^2 overall goodness-of-fit and ANOVA F for row effects: 2×2×2 table, 10 observations per cell, $p = 0.5$ (case 2.3). 128

A1.16 Conditional G^2 backward and ANOVA F with arcsine transformation testing for row effects: 2×2×2 table, 10 observations per cell, $p = 0.1$ (case 2.1). 129

A1.17 Conditional G^2 backward and ANOVA F with arcsine transformation testing for row effects: 2×2×2 table, 10 observations per cell, $p = 0.25$ (case 2.2). 130

A1.18 Conditional G^2 backward and ANOVA F with arcsine transformation testing for row effects: 2×2×2 table, 40 observations per cell, $p = 0.25$ (case 2.2). 131

A1.19 Conditional G^2 backward and ANOVA F with arcsine transformation testing for row effects: 2×2×2 table, 10 observations per cell, $p = 0.5$ (case 2.3). 132

A1.20 Conditional G^2 backward and ANOVA F with arcsine transformation testing for row effects: 2×2×2 table, 40 observations per cell, $p = 0.5$ (case 2.3). 133

A1.21 Conditional G^2 backward and ANOVA F for row effects: 2×2 table, 40 observations per cell, $p = 0.1$ (case 1.2). 134

A1.22 Conditional G^2 backward and ANOVA F with arcsine transformation for row effects: 2×2 table, 40 observations per cell, $p = 0.1$ (case 1.2). 135

A1.23 Conditional G^2 forward and ANOVA F with arcsine transformation testing for row effects: 2×2 table, 10 observations per cell, $p = 0.1$ (case 1.2). 136

A1.24 Conditional G^2 forward and ANOVA F with arcsine transformation testing for row effects: 2×2 table, 40 observations per cell, $p = 0.1$ (case 1.2). 137

A1.25 Conditional G^2 forward and ANOVA F with arcsine transformation testing for row effects: 2×2 table, 10 observations per cell, $p = 0.25$ (case 1.4). 138

A1.26 Conditional G^2 forward and ANOVA F with arcsine transformation testing for row effects: 2×2 table, 40 observations per cell, $p = 0.25$ (case 1.4). 139

A1.27 Conditional G^2 forward and ANOVA F with arcsine transformation testing for row effects: 2×2 table, 10 observations per cell, $p = 0.5$ (case 1.6). 140

A1.28 Conditional G^2 forward and ANOVA F with arcsine transformation testing for row effects: 2×2 table, 40 observations per cell, $p = 0.5$ (case 1.6). 141

A1.29 Conditional G^2 backward and ANOVA F for row effects: 2×2 table, 10 observations per cell, $p = 0.25$ (case 1.4). 142

A1.30 Conditional G^2 backward and ANOVA F with arcsine transformation testing for row effects: 2×2 table, 10 observations per cell, $p = 0.25$ (case 1.4). 143

A1.31 Conditional G^2 backward and ANOVA F for row effects: 2×2 table, 40 observations per cell, $p = 0.25$ (case 1.4). 144

A1.32 Conditional G^2 backward and ANOVA F with arcsine transformation testing for row effects: 2×2 table, 40 observations per cell, $p = 0.25$ (case 1.4). 145

A1.33 Conditional G^2 backward and ANOVA F for row effects: 2×2 table, 10 observations per cell, $p = 0.5$ (case 1.6). 146

A1.34 Conditional G^2 backward and ANOVA F with arcsine transformation testing for row effects: 2×2 table, 10 observations per cell, $p = 0.5$ (case 1.6). 147

A1.35 Conditional G^2 backward and ANOVA F for row effects: 2×2 table, 40 observations per cell, $p = 0.5$ (case 1.6). 148

A1.36 Conditional G^2 backward and ANOVA F with arcsine transformation testing for row effects: 2×2 table, 40 observations per cell, $p = 0.5$ (case 1.6). 149

Chapter 1

Introduction

When all variables are categorical (*i.e.*, discrete and unordered) and observations are taken in the form of a contingency table (or cross classification) loglinear analysis is generally assumed to be the most appropriate method for modelling the observed relationships between the variables. Although they involve many strong parallels to other linear models, loglinear models are different from most common linear models in that they do not identify one variable as the dependent variable.

Logit models, however, provide a means of modelling dependency among categorical variables. For the completely categorical case, loglinear models may be used to fit equivalent logit models, the relationship being that a logit model requires that the marginal totals of the response variable be fixed in the corresponding loglinear model. Usually the response variable is dichotomous so that the logit formulation can be in terms of simple odds of the two categories. The same data analyzed by logit models where the response variable is dichotomous may be viewed as outcomes in a series of Bernoulli trials and arranged as responses in a crossed ANOVA layout. The F-test in ANOVA has been shown to be remarkably robust to many violations of the basic assumptions underlying ANOVA models. In the case of Bernoulli type data, it has been observed that fixed-effects, balanced ANOVA is robust when the treatment group probabilities of success (or failure) are not extreme and the sample size is not too small.

The purpose of this thesis is to compare and contrast classical fixed-effects, balanced ANOVA and loglinear analysis (logit models) as competing methods for analyzing Bernoulli-type data. There is a fundamental difference in the two for this type of data. ANOVA models describe the dependency in terms of additive probability effects whereas logit models describe multiplicative effects for odds that become additive in the logarithm. There is a direct relationship between main effects for the two types of models, but the correspondence ceases for first and higher order interaction effects. Therefore, interaction effects may be present in one model while absent in the other. The focus of investigation in this study is the relative performance of the two types of modelling when the presence of effects is the same for both interpretations.

An assortment of 2×2 and $2\times2\times2$ tables (ANOVA layout perspective) that vary in the group treatment probabilities of success are examined via Monte Carlo simulations. The simulations primarily investigate completely null models and models with main effects present for both types of modelling. The comparison is based mostly upon the Type I error levels and power results of the F-test for ANOVA and the conditional likelihood ratio test G^2 for loglinear fits. Small and moderate sample sizes are used along with a full range of probabilities. In addition, results are obtained to examine the effects of using the Freeman-Tukey arcsine transformation [FrT50] in ANOVA and the effects of adding 0.5 to all elementary cell totals for loglinear analysis. The overall goodness-of-fit statistics for loglinear modelling, the Pearson X^2 and the likelihood ratio G^2, are compared. Though these statistics have already been carefully examined in other studies, the results in this thesis are reported since the sampling scheme, product multinomial, is different from the more commonly investigated full multinomial sampling. Also, stepwise forward and stepwise backward model selection strategies are looked at for the loglinear modelling

Brief reviews of ANOVA and loglinear analysis are given in Chapter 2 and in Chapter 3, respectively. These are quite general and elementary so those already familiar with the methodologies may skip over them without any loss of continuity (except for sections 3.5 and 3.6 of Chapter 3 where logit models and the analogy with ANOVA models are discussed). Chapter 4 contains a review of some of the previous studies and papers related to this thesis. The Monte Carlo simulation design is laid out in Chapter 5. In Chapter 6 the results of the study are presented and discussed. Chapter 7 is the conclusion providing a concise summary of the results.

Chapter 2

Analysis of Variance

This chapter presents a brief overview of the statistical procedure known as the analysis of variance, often abbreviated ANOVA. ANOVA is a hypothesis-testing method enabling one to simultaneously compare the means of several different samples and decide whether an observed difference in the means is the result of a difference in the population means or the result of sampling error alone.

A result of the work of R.A. Fisher (*e.g.*, see [Fis44]), ANOVA has been an invaluable statistical tool for researchers in the biological and social sciences for more than half a century. Accordingly, the literature on the subject of ANOVA combined with modern principles of experimental design is extensive and complete with many monographs and texts having been written on the subject. The reader who desires an detailed and comprehensive look at ANOVA, experimental design, and related issues is referred to any of the following texts: [Lin53], [Sch59], [DuC87], [Kir82]. The aim of this chapter is to review the basic ideas of ANOVA and to introduce the issues most pertinent to this thesis. Those already familiar with ANOVA may wish to skim this chapter, though a quick review of the assumptions underlying the ANOVA (Section 2.3) may be of interest since several are violated by the designs looked at in this thesis.

2.1. The Basics of ANOVA

ANOVA is most often used to analyze tabulated data from an experimental design. The type of ANOVA applied to the data depends upon the design of the experiment. Since there are a great number of available, effective designs from which the experimenter may choose, the number of procedurally different ANOVAs is also quite large. However, the basic rationale behind ANOVA is the same regardless of which design is used by the researcher.

Because of its relative simplicity, the one-way analysis of variance will be used here to present the fundamental ideas of ANOVA. The one-way ANOVA is used to examine $K > 1$ groups (sampled populations) that are assumed to be equivalent in every way except for the treatment applied to each group. If the treatments create any statistically significant difference between the means of each group, then the treatments are said to have different effects on their respective groups. Table 2.1 shows a general layout for the data in a one-way design.

FACTOR 1					
T1	T2	T3	T4	...	TK
Y_{11}	Y_{21}	Y_{31}	Y_{41}	...	Y_{K1}
Y_{12}	Y_{22}	Y_{32}	Y_{42}	...	Y_{K2}
Y_{13}	Y_{23}	Y_{33}	Y_{43}	...	Y_{K3}
Y_{14}	Y_{24}	Y_{34}	Y_{44}	...	Y_{K4}
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
Y_{1n}	Y_{2n}	Y_{3n}	Y_{4n}	...	Y_{Kn}
$\bar{Y}_{1+}$	$\bar{Y}_{2+}$	$\bar{Y}_{3+}$	$\bar{Y}_{4+}$	...	$\bar{Y}_{K+}$

Table 2.1 General layout for one-way ANOVA design.

In the above layout the K treatments (T1,T2,...,TK) comprise K levels of factor 1. This is to say that the treatments are varying degrees or qualities of a common factor. For example, the factor may be temperature with treatments cold, cool, mild, warm, hot. The Y_{jl} (j=1,...,K ; l=1,...,n) are individual observations. In the temperature example the observations may be, say, the lifespan of a specific manufacturer's battery type measured in hours. Table 2.1 shows n observations per treatment group. The final row of $\bar{Y}_{j+}$ are the sample means of each treatment group. The subscript indicates summation over the index replaced by the addition sign.

The question that the experimenter seeks to answer is: Do the various treatments have significantly different effects on the observed values? Examination and comparison of the group means is one approach to answering the question. However, even in the case where all the treatments have no effects, one would expect some variation among the group means due to sampling error. Therefore, mere inspection of means is not sufficient to answer the question since one does not generally know how much variation is attributable to chance and random error alone in any given experiment (or sample survey). Analysis of variance is a hypothesis test that compares the variability between all of the sample means with what one would expect by chance if the population means were equal. The null hypothesis in ANOVA is that the means of the populations from which the samples were drawn are equal. This is the same as assuming that there are no differences in treatment effects.

The logic underlying ANOVA for testing the null hypothesis is to calculate two different estimates of the population variance, σ^2, which is assumed homogeneous within groups. The first is referred to as the within-groups estimate of σ^2, or $\hat{\sigma}^2_{WG}$. The second is referred to as the between-groups estimate of $\sigma^2_{\bar{Y}}$, or $\hat{\sigma}^2_{BG}$ and is based on the group means. The ratio $\hat{\sigma}^2_{BG}/\hat{\sigma}^2_{WG}$ is the statistic used to test the hypothesis. Under the null hypothesis of no treatment effects (*i.e.*, equal group population means) the ratio is distributed as a central F distribution with ν_1 and ν_2 degrees of freedom in the numerator and denominator, respectively. Here ν_1 is the number of groups minus one and ν_2 is the total number of observations minus the number of groups. In this case, one is interested in the variance of the sample means of the individual groups. If the null hypothesis of no treatment effects is true, then the observations in each group should all belong to the same parent population. Therefore, under the null hypothesis, the variance of each treatment group is an unbiased estimator of the population variance. Since there are K such estimators the average of these estimates provides a better estimate of the population variance than a choice of any one of the individual group estimates (sufficiency upheld). This average of the group estimates of σ^2 is the within-groups variance, $\hat{\sigma}^2_{WG}$, ("within-groups" indicates that the variance is estimated from the variances of the observations within each treatment group) and can then be used to estimate the variance of the sampling distribution of the mean, $\sigma^2_{\bar{Y}}$. It is well known that $\sigma^2_{\bar{Y}}$ equals σ^2/n, where n is the size of each group.

The within-groups estimate of $\sigma^2_{\bar{Y}}$ is unaffected by the presence of group treatment effects. To see why this is so recall that adding a constant to each member of a distribution increases the mean of the group but not the variance. Since the within-groups estimate of $\sigma^2_{\bar{Y}}$ is the same independent of the presence or lack of presence of treatment effects, calculation of another estimate of $\sigma^2_{\bar{Y}}$ that is unbiased only when the null hypothesis of no treatment effects is true leads to the desired test statistic.

The between-groups estimate of $\sigma_{\bar{Y}}^2$ is such an estimate. It is found by viewing the group means as independent estimates of the population mean and then computing the variance of that sample. If the treatments have no effects on the group means then the $\hat{\sigma}_{BG}^2$ is an unbiased estimator of $\sigma_{\bar{Y}}^2$. If any significant treatment effects do exist then $\hat{\sigma}_{BG}^2$ will be larger than $\sigma_{\bar{Y}}^2$ and larger than $\hat{\sigma}_{WG}^2$.

The remaining problem is to resolve what constitutes a significant (beyond chance alone) difference between $\hat{\sigma}_{BG}^2$ and $\hat{\sigma}_{WG}^2$. Fortunately, the probability density function[1] for the F-ratio which describes the distribution of $\hat{\sigma}_{BG}^2/\hat{\sigma}_{WG}^2$ when the null hypothesis of no treatment effects is true has been determined. Hence, one can compare $\hat{\sigma}_{BG}^2/\hat{\sigma}_{WG}^2$ with the tabulated critical values of the appropriate F distribution to test the null hypothesis.

2.2. Fixed-Effects Two-Way and Three-way ANOVA

The previous section gave a simplified, intuitive introduction to the one-way analysis of variance. The ideas are readily extended to more complicated experimental designs. Particular to this thesis are the fixed-effects designs for two and three-way ANOVA. These are the models that were chosen for comparing to the analogous log-linear models that will be presented in the next chapter.

The two and three-way designs are straightforward extensions of the one-way design. The difference is in the inclusion of another factor (with corresponding treatments) for the two-way design, and the inclusion of two additional factors for the three-way design. The inclusion of additional factors creates a cross-classification. This means that in a two-way design each observation is subjected to two treatments, one from each factor. In the two-way designs these treatments are often labelled row and column treatments. The effects of these treatments are referred to as main effects. This is due to the fact that each effect is attributable to either a row or column factor, but not to both at the same time.

---

[1] The density function depends upon the degrees of freedom in the numerator and the denominator. There is a *family* of F distributions.

		FACTOR J	
		T1	T2
FACTOR I	T1	$Y_{111},\ldots,Y_{11n}$	$Y_{121},\ldots,Y_{12n}$
	T2	$Y_{211},\ldots,Y_{21n}$	$Y_{22n},\ldots,Y_{22n}$

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}; \qquad i = 1,2;\ j = 1,2;\ k = 1, \ldots, n;$$

μ = population mean;
α_i = effect of treatment i of factor I;
β_j = effect of treatment j of factor J;
$(\alpha\beta)_{ij}$ = effect of interaction of treatments i and j of factors I and J;
ε_{ijk} = error component of observation Y_{ijk};

Table 2.2 Table layout and parametric equation for 2×2 ANOVA examined in this study (T1 and T2 denote treatments).

In addition to possible row and column effects in the two-way designs there is also the possibility of interaction effects. Interaction effects represent the deviation of observations from the overall mean that are not accounted for by the main effects, yet are still significant beyond chance alone. They are the result of a combination of treatments. Intuitively, a combination of treatments may result in an effect above and beyond the individual treatment effects as is the case when a person mixes various medications with various amount of alcohol consumption. In the case of an interaction effect the treatments have a catalytic or synergistic influence.

The hypotheses tested in the two-way and three-way designs are ones concerned with the presence or absence of the possible effects. Therefore, in the two-way design, three null hypotheses may be tested : no row effects, no column effects, no interaction effects. The usual tests of these hypotheses are well known to be independent. For each a separate F-ratio is calculated and tested for significance. Likewise, for three-way and higher designs, hypothesis tests can be independently carried out for each possible effect. The interaction effects are usually referred to by the number of factors involved in the interaction. For example, the interaction effects in a three-way design are the two-factor interactions and the three-factor interactions. Table 2.2 and Table 2.3 depict the layouts and corresponding notations used in this study.

		FACTOR K			
		T1		T2	
		FACTOR J		FACTOR J	
		T1	T2	T1	T2
FACTOR I	T1	$Y_{1111},\ldots,Y_{111n}$	$Y_{1211},\ldots,Y_{121n}$	$Y_{1121},\ldots,Y_{112n}$	$Y_{1221},\ldots,Y_{122n}$
	T2	$Y_{2111},\ldots,Y_{211n}$	$Y_{2211},\ldots,Y_{221n}$	$Y_{2121},\ldots,Y_{212n}$	$Y_{2221},\ldots,Y_{222n}$

$$Y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\alpha)_{jk} + (\alpha\beta\gamma)_{ijk} + \varepsilon_{ijkl};$$

$$i = 1,2;\ j = 1,2;\ k = 1,2;\ l = 1,\ldots,n;$$

μ = population mean;
α_i = effect of treatment i of factor I;
β_j = effect of treatment j of factor J;
$(\alpha\beta)_{ij}$ = effect of interaction of treatments i and j of factors I and J;
$(\alpha\gamma)_{ik}$ = effect of interaction of treatments i and k of factors I and K;
$(\beta\gamma)_{jk}$ = effect of interaction of treatments j and k of factors J and K;
$(\alpha\beta\gamma)_{ijk}$ = effect of interaction of treatments i and j and k of factor I and J and K;
ε_{ijkl} = error component of observation Y_{ijkl};

Table 2.3 Table layout and parametric model for 2×2×2 ANOVA examined in this study (T1 and T2 denote treatments).

2.3. Assumptions Underlying ANOVA F-Test

Several important assumptions have to be made and considered when using the F-test of the null hypothesis in the ANOVA discussed so far. These assumptions are particularly important in this thesis since they will be subjected to violations. Discussion of these violations will be reserved until Chapter 4. At the moment the discussion will be limited to introducing the assumptions in general. The assumptions are best expressed in terms of the parametric model that underlies the chosen design. The models for the two-way and the three-way designs are presented in Table 2.2 and in Table 2.3, respectively. Briefly, the assumptions are as follows :

[1] Before administration of the treatments, all samples are drawn at random from the same population[2].

---

[2] The assumption of same population may be relaxed [Lin53p. 51] but the inferences from the hypothesis test must be modified. The point of ANOVA is to check for equal treatment effects. The F-ratio is still distributed as F even if the treatment samples come from different treatment population before application of treatments. However, if this is the case, applications of treatments could conceivably result in equal treatment group means (when originally they were different) which would mean treatment effects are present. Yet, in such a case the F-test would not lead to rejection of the null hypothesis; a

After the treatments have been applied, the sampled populations (there are ij for the two-way and ijk for the three-way designs) may be viewed as random samples from corresponding crossed treatment populations. This assumption effectively requires that the deviations from the population means be statistically independent both within and across treatment combinations. This leads to the related assumption that the samples are distributed normally with mean zero and variance equal to the error variance, $N(0,\sigma_\varepsilon^2)$.

[2] The variances of the crossed treatment populations are all equal (homoscedasticity).

[3] The distribution of the after treatment scores (the Y_{ijk} in Table 2.2) in each crossed treatment group is normal.

[4] The mean of the scores in each crossed treatment population is the same (the null hypothesis).

With these assumptions it can be shown (using the two-way model, for example, with $i=1,\ldots,a$; $j=1,\ldots,b$) that any set of the $a \times b$ population means can be expressed with $\mu+\alpha_i+\beta_j+(\alpha\beta)_{ij}$ such that

$$\sum_{i=1}^{a}\alpha_i=0\,,$$
$$\sum_{j=1}^{b}\beta_j=0\,,$$
$$\sum_{i=1}^{a}(\alpha\beta)_{ij}=0\,,$$
$$\sum_{j=1}^{b}(\alpha\beta)_{ij}=0\,,$$
$$\sum_{i=1}^{a}\sum_{j=1}^{b}(\alpha\beta)_{ij}=0\,.$$

These constraints are consistent with the property that for any population the sum of all the deviations from the mean of the population must equal zero.[3]

Type II error would result. Therefore, the assumption is best kept if detection of no effects is the hypothesis.

[3]

$$\sum_{i=1}^{n}(x_i-\mu)=-n\mu+\sum_{i=1}^{n}x_i=-n\sum_{i=1}^{n}x_i/n+\sum_{i=1}^{n}x_i=0.$$

The first assumption is perhaps the most important with respect to the inferences made from the ANOVA. Simply put, it requires that all the treatment subjects are in no way statistically related to each other either within groups or across groups. If there is correlation among the subjects that the experimenter has overlooked or ignored, the estimators could show a bias that the experimenter would then incorrectly attribute to the presence of effects. Thus, an experimenter needs to be careful that the assumption is met if the inferences drawn from the ANOVA are to be trusted.

The second assumption is fundamental to the rational underlying the F-ratio in ANOVA. The within-groups variance estimator is automatically biased if there is heterogeneity of the crossed treatment population variances. This is because the test assumes that the sample variance of each group is an estimator of a population variance common to all the groups. Fortunately, the F-test is robust for moderate degrees of heterogeneity. (More will be said about this in Chapter 4).

The final two assumptions are of lesser consequence than the first two. The assumption that the response variables, the Y_{ijk} in the two-way example, are normally distributed can generally be relaxed to one of assuming that the forms of parent distributions of the crossed treatment responses are all similar. The final assumption of equal means of crossed treatment groups is merely the null hypothesis. The F-ratio is only distributed as a central F distribution when the null hypothesis is true. This then is what provides the test to accept or reject the null hypothesis.

2.4. Model Selection

The F-test for the possible effects may be made individually and independently of each other, but usually are calculated at the same time. When a test turns up a significant result, the corresponding effect parameter is added to the model. If the experimenter's only concern is to test for a specific effect then the model is of little importance and not necessary. On the other hand, if the experimenter seeks a model to express the response variable in terms of the independent variables (factors), then all effects must be investigated for significance.

The different significant effects can be estimated from the appropriate means. For example, if a row factor is found to be significant in a two-way design, the value of the effect of the first treatment can be estimated by subtracting the mean of the first row from the over-all mean. The other effects may be estimated in a similar way. One needs only the means and the results of the significance tests to make the calculations.

Chapter 3

The Loglinear Model

This chapter gives a basic introduction of the loglinear model for categorical data. For simplicity most of the discussion is limited to the 2 dimensional model since generalizations to higher dimensions are straightforward and do not require any adjustments or additions to the basic theory. However, it should be stated at the outset that the main value of loglinear analysis is found in its applications for analyzing higher dimensional cross-classifications and this is the purpose for which it was developed. After presentation of the general loglinear model the logit model is introduced. The logit model is shown to be a restricted case of the general loglinear model. This distinction is made since it is the logit model (via its general loglinear representation) that is being compared with ANOVA in this thesis.

3.1. History

Loglinear analysis evolved from the need for a general analytical tool for studying multi-way cross-classifications of categorical data. The theory surrounding two-way tables has been well established since the early part of this century (*e.g.*, see [Pea00] and [Yul00]). However, a suitable method for analyzing higher-way tables of categorical data, especially the high order interactions, analogous to that allowed by multiple regression analysis did not begin to appear until the mid-1950s. This was primarily fundamental work on measures of association and methods of estimation [GoK54], [GoK59], [GoK63], [RoK56], [Lan51]. From this emerged the present theory of loglinear analysis, mostly in the 1960s. By the early 1970s several comprehensive overviews of the theory and the applications surrounding loglinear analysis had been published (*e.g.*, [Goo78], [BFH75], [Pla74], [Hab74], [Fie77]).

3.2. The Model

Assume that one has a 2×2 table of counts with underlying cell probabilities, p_{ij}, shown below where $\sum_{ij} p_{ij} = 1$. In this case the row and column variables have $r=2$ and $c=2$ levels, respectively.

p_{11}	p_{12}
p_{21}	p_{22}

Table 3.1 Example 2×2 table of probabilities.

The most general loglinear model for this table is then:

$$log\,(p_{ij}) = u + u_{1(i)} + u_{2(j)} + u_{12(ij)}\,;\ i = 1,2;\ j = 1,2;$$

with the constraints

$$\sum_i u_{1(i)} = \sum_j u_{2(j)} = \sum_i u_{12(ij)} = \sum_j u_{12(ij)} = 0$$

In analogy with ANOVA, the model parameters may be expressed in terms of the cell probabilities as

$$u = \frac{1}{r \cdot c} \sum_{ij} log\,(p_{ij})$$

$$u_{1(i)} = \frac{1}{c} \sum_j log\,(p_{ij}) - u$$

$$u_{2(j)} = \frac{1}{r} \sum_i log\,(p_{ij}) - u$$

$$u_{12(ij)} = log\,(p_{ij}) - u - u_{1(i)} - u_{2(j)}$$

Hence, u is the grand mean of the logarithms of the probabilities, $u_{1(i)}$ and $u_{2(j)}$ are the main effects (or deviations from the grand mean) for levels i and j of variables 1 and 2, respectively, and $u_{12(ij)}$ is the interaction effect for cell ij.

The above model is called saturated since it includes all possible u-terms corresponding to all variables and possible interactions. In this case, the estimated counts equal the observed counts since there are as many parameters as there are observations.[4] Here $u_{12(ij)}$ is called the highest-order u-term since it represents interaction of all variables in the model. The u-terms whose subscripts are subsets of the subscripts of other higher-order terms are called lower-order relatives. A loglinear model is hierarchical if it satisfies the following two conditions:

---

[4] The model can be rewritten in terms of expected counts. Since expected count $m_{ij} = Np_{ij}$ where N is the grand total of counts, the saturated model may be written $log\,(m_{ij}) = log\,(N) + log\,(p_{ij}) = log\,(N) + u + u_{1(i)} + u_{2(j)} + u_{12(ij)}$.

[1] If a u-term is set equal to zero, then all higher-order relatives of that term are zero.

[2] If a u-term is not equal to zero then all of its lower-order relatives are not zero.

Here "u-term equal to zero" means that the term is zero for all levels or combinations of levels (*e.g.*, $u_{12(ij)} = 0$ for all i, j).

Only hierarchical models are examined in this study. Hierarchical models are desirable since they permit easy and straightforward calculation of cell estimates. This is not true for nonhierarchical models, in which case, for the same maximum likelihood procedures to apply, the tables need to first be transformed so that a hierarchical model can be formed. Also, if higher-order effects are to be interpreted as measuring deviations from their lower-order relatives the hierarchical assumption is needed.

Still, there do exist cases where nonhierarchical models are the most appropriate to describe the data. These situations arise when there exists a synergistic relationship between two or more variables. That is, there is an effect when two or more variables are together but no effect exists when each variable is taken alone. Consequently, interpretation of the higher-order u-terms included in nonhierarchical models is awkward if not difficult. Therefore, use of hierarchical models is considerably more popular in applications. More complete discussions of hierarchical models and references to aspects of nonhierarchical loglinear modelling may be found in [Fie77] and [Hab74].

The loglinear formulation discussed so far is a linear model in the natural logarithms of the cell probabilities. This linearity is not intrinsic to building a model for expressing the expected cell probabilities. Rather, it is a feature that allows for parameter interpretations directly analogous to those of other common linear models such as those in ANOVA and multiple regression. For each loglinear model there is an underlying multiplicative model for explaining the data. The multiplicative models become linear in their logarithms thereby giving the corresponding loglinear models (see [Ken83] for an introduction to loglinear analysis with attention given to multiplicative formulations). One may argue that the multiplicative version better reflects the maximum likelihood estimation used in deriving cell expectations (*i.e.*, the maximum likelihood cell estimates are ratios of products of the observed marginal totals). Nevertheless, general preference has been given to the loglinear formulation in both the applied and theoretical literature. In this study it is natural to select the loglinear model for making the comparisons with ANOVA since these two types of models have analogous interpretations.

3.3. Sampling Schemes

The three most common sampling schemes for gathering cross-classified data are:

[1] *Full multinomial sampling* : A fixed random sample of size N is taken and each observation is independently selected and then allocated to a cell (*i.e.*, the cell count is incremented by 1) based on the agreement of its attributes to the categories of the variables in the design. Each observation must be allocated to only one cell of the table and the variable categories must be exhaustive so that every observation can be categorized within the table.

[2] *Product multinomial sampling* : The marginal totals are fixed for a variable or combination of variables, often considered explanatory variables. These marginal totals are those found by summing over the levels of the response variable(s). For example, given an $I \times J \times K$ table where the response variable is indexed by $k = 1, \cdots, K$ product multinomial sampling yields $I \times J$ independently sampled multinomials each with K categories. The sizes of each of the $I \times J$ multinomials do not have to be the same.

[3] *Poisson sampling* : Each cell count is assumed to have an independent Poisson distribution. Neither the grand total nor any of the marginal totals are fixed. Instead, observations are made over a fixed period of time and the totals are recorded for each cell.

Loglinear modelling is appropriate for each of these sampling schemes. They all lead to the same model estimates and thus the same goodness-of-fit statistics. However, in the case of product multinomial sampling, the hypothesized model must include the u-terms corresponding to the fixed margins. [Bir63]. Product multinomial sampling is the scheme chosen for this study since it corresponds to the models where one variable is designated as the response variable, thereby continuing the analogy with ANOVA. Such models form the set of logit models. Logit models will be discussed in Section 3.5.

3.4. Estimation and Model Selection

Once a loglinear model has been proposed for the data, estimates of cell counts under the model can be found using the method of maximum likelihood [Bir63]. The maximum likelihood estimates are functions of the observed marginal totals of the table. The model specifies which marginal totals are minimally sufficient for generating the estimates. They are the marginal totals that correspond to u-terms in the models which have no higher-order relatives. Take, for instance, the 3 dimensional model

$$log(m_{ijk}) = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)}$$

The minimally sufficient marginal totals are x_{++3} and x_{12+} which correspond to the terms $u_{3(k)}$ and $u_{12(ij)}$, respectively. The addition sign in the subscript indicates summation over the categories of the variable it replaces.

With these totals the estimates may be calculated given the requirements of Birch that:

[1] the minimally sufficient marginal totals of the estimates must equal the minimally sufficient marginal totals of the observations, and,

[2] that the set of estimates satisfying those equalities and model constraints is unique. There is a problem, however, when these solutions give cell estimates equal to zero, which will occur when there is a zero in any of the minimally sufficient marginal totals of the observations. In those situations maximum likelihood estimates do not exist for the u-terms corresponding to those minimally sufficient marginal totals. Some useful suggestions for dealing with this problem are outlined in [Fie77]. Other

forms of estimation such as weighted least squares [GSK69] also break down when there are many observed zero counts.

Assuming there are no zeros in the minimally sufficient marginal totals, estimates may be found by either directly solving for a closed form solution that satisfies the model and maximum likelihood constraints or by using an iterative method. The two most common iterative methods are the Newton-Raphson techniques (see [Hab74], [Hab78], and [Hab79] for applications to loglinear analysis) and iterative proportional fitting [DeS40]. Both iterative methods provide estimates when closed form solutions do not exist (see [BFH75] for a method of detecting when they do) and both provide the same estimates as those found directly using the closed form solutions.

After cell estimates have been obtained for a model, the fit of the model to the observed data may be subjected to significance testing by calculating the Pearson chi-square statistic

$$X^2 = \sum_{allcells} \frac{(observed - expected)^2}{expected}$$

or the likelihood-ratio statistic

$$G^2 = 2 \sum_{allcells} (observed) \log \frac{observed}{expected}$$

Both statistics have asymptotic χ^2 distributions with v degrees of freedom where $v = \#cells - \#parameters\ in\ hypothesized\ model$. These statistics measure the overall goodness-of-fit of the model. Significant values give reason to reject the model and search for an alternative.

This brings up the topic of model selection. The object of model selection is to find the most parsimonious model (*i.e.*, most restrictive in that it uses fewer parameters) that still adequately fits the observed data. One could calculate G^2 and X^2 for every possible model and then select the one which is most parsimonious and has least significant statistics. However, more often than not there will be several "best fits" using this procedure and the experimenter may not be able to choose one from among them when the differences are slight or the choice runs contrary to the experimenter's insight.

A more common method of model selection involves using a property of G^2 that is not shared by X^2. The G^2 statistic can be partitioned to test nested hypotheses. Nested hypotheses arise when all the parameters of one model are included in another model— *i.e.*, where a more general model is being compared with one with a reduced number of parameters. Then the hypothesis that the nested, more parsimonious model is acceptable conditional to the assumption that the larger model is true can be tested. The difference in the values of the G^2 statistics, G^2 of smaller model minus G^2 of larger model, is approximately distributed as χ^2 with degrees of freedom equal to the difference of the degrees of freedom for the two models. This statistic is often referred to as the conditional likelihood ratio statistic. It is denoted usually as $G^2_{A|B}$ where the subscript *A* stands for the nested model and *B* stands for the larger model (*i.e.*, model *A* given that model *B* is true). $G^2_{A|B}$ can also be expressed as

$$G^2_{A|B} = 2 \sum_{\text{all cells}} [expected]_B \log \frac{[expected]_B}{[expected]_A}$$

Hence, the cell count estimates calculated for model *B* are used like observed values in the regular G^2.

Given this property of G^2, stepwise selection strategies have become popular. These are stepwise forward selection and stepwise backward selection. Use of these techniques involves incrementing or decrementing the number of *u*-terms in the model and examining the conditional G^2 statistic. For stepwise forward selection all *u*-terms of the next higher order in the hierarchy of models are added and tested for significance one at a time. The most significant of these terms is appended to make a new base model and the process is repeated for the remaining *u*-terms of that order. Should all the *u*-terms of the same order be added to the base model, the *u*-terms of the next higher order are tested. Stepwise backward selection is similar except the starting model is the saturated model and the process successively deletes non-significant terms from the model. These procedures may also be mixed. When the cross-classification has a large number of variables the overall G^2 and X^2 statistics may be examined first for full order models to see where to begin. A full order model is one that satisfies the condition that if a *u*-term is in the model, then all the *u*-terms of that order are also included in the model.

Often the aforementioned procedures still do not identify a single model as the best fit in which case some follow-up analysis can be carried out. Individual effects parameters may be standardized and examined for significance. Standardized cell residuals may also be considered. The two most common methods for this type of follow-up involve examination of either components of X^2 or Freeman-Tukey deviates. References and descriptions of these follow-up procedures, and more, can be found in the texts cited at the end of section 3.1 in this chapter.

3.5. Logit Model

The logit model is of specific interest in this thesis since it and the fixed effects, balanced ANOVA model are used to analyze the same sets of data in the Monte Carlo study outlined in Chapter 5. Logit models are a subset of loglinear models when all data is categorical. In an asymmetrical study of the variables one or several of the variables in the cross-classification are assumed to be explanatory while the others are response variables. The words response and explanatory may be interchanged with dependent and independent, respectively. The logit model posits a dependency relationship between the response variable(s) and the explanatory variable(s). For example, suppose one choses to investigate whether or not vote(yes, no) depends on income(high, middle, low), gender(female, male), and education(univ, high school, elementary). The response variable is vote and the explanatory variables are income, gender, and education.

Discussion is confined to the cases where there is only one response variable and two or three explanatory variables. The study of multiple response variables involves structural equations models which is a large part of the topic referred to as path analysis (see [Dun75] for a detailed treatment). Multiple categories in the response variable present no problems for logit analysis [BFH75], but they do create a problem if the analogy with ANOVA is to be maintained. Therefore, it is also assumed that the response variable has only two categories. This forms the analogy with ANOVA where the response is the outcome of a Bernoulli trial. Hence, the two categories of the response variable in the analogous logit formulation would be, say, *success* and *failure*. With the restrictions just mentioned the parameters in the logit model have a direct correspondence to the parameters in the fixed effects ANOVA model (except logit models and loglinear models have no error term).

With a two category response variable the logit model describes the behavior of the logarithm of the odds of one category of the response variable to the other in terms of the effects of the explanatory variables. Assume that there is a 3 dimensional model with variables 1, 2, and 3 where variable 3 is the dichotomous response variable. The logit model analogous to the ANOVA model with both main effects and interaction effects is

$$logit(i,j) = \log\frac{m_{ij1}}{m_{ij2}} = w^{3}+w^{13}_{(i)}+w^{23}_{(j)}+w^{123}_{(ij)}$$

where the w-terms sum to zero over any of their indices. The relationship of the logit model to the loglinear model is straightforward:

$$\begin{aligned}
\log\frac{m_{ij1}}{m_{ij2}} &= log(m_{ij1}) - log(m_{ij2}) \\
&= [u+u_{1(i)}+u_{2(j)}+u_{3(1)}+u_{12(ij)}+u_{13(i1)}+u_{23(j1)}+u_{123(ij1)}] \\
&\quad -[u+u_{1(i)}+u_{2(j)}+u_{3(2)}+u_{12(ij)}+u_{13(i2)}+u_{23(j2)}+u_{123(ij2)}] \\
&= [u_{3(1)}-u_{3(2)}]+[u_{13(i1)}-u_{13(i2)}]+[u_{23(j1)}-u_{23(j2)}]+[u_{123(ij1)}-u_{123(ij2)}] \\
&= 2\,[u_{3(1)}+u_{13(i1)}+u_{23(j1)}+u_{123(ij1)}] \\
&= w^{3}+w^{13}_{(i)}+w^{23}_{(j)}+w^{123}_{(ij)}
\end{aligned}$$

The second to last equality is due to the fact that variable 3 has two levels and the u-terms are constrained to sum to zero across levels. The u-terms that have the response variable in their subscript in the loglinear model do not cancel out in the transformation to the logit representation. On the other hand, all other u-terms do cancel. This means there could be several loglinear representations for the same logit model. The reason for this is that the logit model does not put any restrictions on the relationships between the dependent variables. However, arguments have been made (*e.g.*, [BFH75]) that the loglinear model should contain the u-term for highest order interaction of the dependent variables (implying it contains all lower order relatives due to the hierarchy principle). The main defense for doing this is that if the analysis is to explain the response in terms of the explanatory variables, then the tests for individual effects parameters (using conditional G^2) should partial out the relationships not involving the response variable. Including all u-terms involving the explanatory variables in the two models being compared by the conditional G^2 test removes the influence of the relationships of the explanatory variables from the tests. This is analogous to regression analysis using generalized least squares estimation (not ordinary least squares estimation) where correlations among independent variables are accounted for by transforming the data (using the information in the variance-covariance matrix) so that the error terms are uncorrelated. In this study all u-terms involving the explanatory variables are included in the loglinear formulations of the logit models.

3.6. ANOVA/Loglinear Analogy—Example

To conclude this chapter a hypothetical example depicting an ANOVA model and the corresponding loglinear model where both are used to describe the same set of data is presented. The example extends to all the comparisons made in the Monte Carlo study described in Chapter 5. Consider a 2×2 table of Bernoulli trial success counts where there are 10 trials per cell (Table 3.2). The ANOVA model for this table has two explanatory variables each with two categories. The response variable is the outcome of the Bernoulli trial, 1 for success, 0 for failure.

	var 3	
var 2	1 1 1 0 0 0 0 0 0 0 $\sum_{all}=3$	1 1 1 1 1 0 0 0 0 0 $\sum_{all}=5$
	0 0 0 0 0 1 1 1 1 1 $\sum_{all}=5$	1 1 1 1 1 1 1 0 0 0 $\sum_{all}=7$

Table 3.2 Example of Bernoulli trial data for 2×2 ANOVA.

The ANOVA model for both main effects present and no interaction effects is

$$Y_{ijk} = \mu+\alpha_j+\beta_k+e_{ijk}$$

The analogous loglinear model is

$$log\,(m_{ijk}) = u+u_{1(i)}+u_{2(j)}+u_{3(k)}+u_{23(jk)}+u_{12(ij)}+u_{13(ik)}$$

where index i is for the two category response variable. Hence, the logit formulation is

$$logit\,(j,k) = \frac{log\,(m1jk)}{log\,(m2jk)} = w^1+w_{(j)}^{12}+w_{(k)}^{13}$$

This is like the example in the previous section except there is no interaction and the response variable has index i instead of index k. This model examines the same data as the ANOVA except it is arranged differently (Table 3.3).

		var 1			
		success		failure	
		var 3		var 3	
		cat 1	cat 2	cat 1	cat 2
var 2	cat 1	3	5	7	5
	cat 3	5	7	5	3

Table 3.3 Example contingency table for data in Table 3.2 (above).

The subtable under the success category of variable 1 is equivalent to the success totals for each cell of the ANOVA data in the Table 3.3. The subtable under the failure categories is the number of failures in each cell of the ANOVA table. Under this sampling scheme the marginal totals found by summing over variable 1 must all equal 10. The sampling is product multinomial type (4 multinomials of overall size 10, each with two categories). Hence, in the ANOVA formulation a scoring of the observations (0 or 1) represents the response whereas in the loglinear formulation the response variable is explicitly categorical. The same data is interpreted differently for the two methods. For ANOVA the data are viewed as real number values but for loglinear analysis the data are viewed as category tallies.

Chapter 4

Related Studies

Though the analogy of the loglinear model to the ANOVA model is made in almost all introductory overviews of loglinear analysis, few comparisons of the two methods for analyzing the same data sets have been made, and those that have been carried out are limited. Undoubtedly, this is due to the very obvious theoretical differences in the interpretations and applications of the two methods. Most conspicuous are the different assumptions about the data that each method is appropriate for analyzing. Loglinear analysis is intended for categorical data whereas ANOVA is intended for continuous data. Extending the use of either method to analyze data that is more appropriately analyzed by the other method involves violations of basic assumptions. The loglinear model is certainly not viable for analyzing continuous data unless the data is transformed by scoring or partitioned into categories and counted. On the other hand, ANOVA techniques are surprisingly robust.

This chapter provides a brief survey of some studies related to the topic of this thesis. First, studies pertaining to loglinear analysis are reviewed. Next, some studies of the robustness of ANOVA are noted. The last section discusses some comparisons that have been made between loglinear analysis and ANOVA.

4.1. Studies Related to Loglinear Analysis

A good deal of attention has been given to the small sample behavior of the goodness-of-fit statistics X^2 and G^2. As these are the primary statistics for testing loglinear models, results concerning them are important in this study. It has been established that both of these statistics vary from there asymptotic distributions when many cell counts are small or zero. This is the situation where most attention has been focussed with respect to the two statistics.

A thorough investigation of small sample behavior of X^2 and G^2 is given in [Lar78]. Larntz obtained both exact and Monte Carlo results in examining Type I error over a variety of multinomials for 1, 2, and 3 dimensional tables. The results demonstrate that the Pearson X^2 statistic out-performs G^2 by a wide margin.

For expected cell counts in the range 0–1.5 G^2 does not reject often enough. For expected counts in the range 1.5–4.0 G^2 rejects much too often. In contrast, X^2 is sufficiently close to nominal when expected counts are greater than 1.0. G^2 becomes acceptable only when expected counts are greater than 4.0. Larntz attributes the high Type I error rates of G^2 when expected counts are moderate to the large contribution to G^2 from vary small observed counts. This conclusion is based upon calculations that were made for both X^2 and G^2 of the exact contributions of various observed counts for several fixed expected counts.

Other studies are in agreement with the findings of Larntz. In the course of these investigations several rules of thumb have been suggested for determination of the minimal expected number of observations per cell that are needed in order to reliably use X^2 and G^2. A concise survey of these rules may be found in [Rud87]. Most of these rules are for the case of a single multinomial without estimated parameters. Loglinear models are considered by [Lar78], [Odo70], and [Rud87]. Fienberg [Fie79] suggests, based on the results of Larntz, that the overall sample size, N, may be as small as 4 or 5 times the number of cells. Rudas carried out Monte Carlo simulations to obtain 90% and 95% confidence intervals for X^2 and G^2 for a variety of 2 and 3 dimensional tables. It was shown that the statistics are reasonably acceptable (*i.e.*, the confidence interval contains the asymptotic percentage point) even for the range of sample sizes equal to 2 or 3 times the number of cells in the table.

These studies have mostly been focussed on examination of the statistics when the null hypothesis is true. Results in [CrG82] are used to examine the power of X^2 and G^2 for multinomials and contingency tables. They use a measure of strength which they define as a weighted average of the power. The results indicate that X^2 and G^2 have about equivalent power, with the power of G^2 being slightly greater.

Other studies involving goodness-of-fit statistics for models of contingency table data have been directed at tailoring special statistics and tests to individual problems. Use of the lognormal approximation of the χ^2 distribution and the use of a scaled χ^2 distribution as reference distributions for the Pearson X^2 statistic are examined in [LaU84]. Their study is confined to only the case of 2 dimensional tables and the chi-squared test for independence, but the results indicate considerable improvement over the traditional chi-squared test for that restricted class. A new goodness-of-fit statistic is introduced in [Sim85] for sparse multinomials that performs well when the assumption that the null distribution exhibits smoothness is met. The smoothness restriction allows for information in neighboring cells to be used collectively to aid in estimating the probabilities for each cell. In an extension of results in [KoL80] Koehler [Koe86] examines G^2 for sparse contingency tables when using the normal approximation in comparison to the usual chi-squared approximation. The normal approximation is shown to be much more accurate than the chi-squared approximation for G^2 in many cases, though possible bias of estimated moments remains a problem for very sparse tables. Several goodness-of-fit statistics for contingency table models based on cluster sampling are compared in [ThR87]. Though not specifically relevant to this thesis, the study is of interest since it extends the sampling schemes and models for contingency table data.

4.2. Studies Related to ANOVA

As ANOVA predates the emergence of loglinear analysis, many studies have been carried out to examine the robustness of ANOVA under violations of basic assumptions or to develop new statistics and transformations to correct for the violations. Of specific interest in this thesis are the results involving ANOVA for discrete data, especially binomially distributed counts. Binomial data, such as that examined in the Monte Carlo study described in Chapter 5, violate several of the important assumptions of the tests used in ANOVA. The error terms and the treatment populations are not normally distributed. In many situations heterogeneity of cell variances will exist since the variances are proportional to the cell means. In addition to this the distributions of the cell populations may be of nonhomogeneous shape even when homogeneity of variance is present.

Early work considering the use of ANOVA for analyzing Bernoulli type data concentrated on transformations of the data to remove heterogeneity of variance and give normally distributed data. Fisher first suggested use of the arcsine transformation [Fis44]. Bartlet and Cochran developed variants of the same transformation [Bar47], [Coc40]. A rigorous development of the mathematical theory pertaining to the transformations was carried out by Curtiss in [Cur43]. Freeman and Tukey [FrT50] introduced a transformation that Mosteller and Youtz showed in [MoY61] to be superior to the other arcsine transformations. Investigations of arcsine transformations apparently subsided after the study by Mosteller and Youtz. No Monte Carlo results comparing ANOVA using transformed data to ANOVA using raw data were found in the literature survey for this thesis. Specific case examples using the arcsine transformation are given in [MoT68] and in [Coc40], but the ANOVA for raw scores is not displayed. Interestingly, when the ANOVA was carried out for the raw data in the Mosteller-Tukey example the results were found to be nearly identical to those for the ANOVA in the example using the transformed scores. (ANOVA for the raw data in the Cochran example was not carried out).

Monte Carlo studies examining ANOVA with binomial populations were carried out by Hsu and Feldt [HsF69] and by Lunney [Lun70]. Hsu and Feldt investigated the robustness of the F test for binomial populations for probabilities of $p = 0.25$, $p = 0.40$, and $p = 0.50$. They examined experiments with 2 or 4 treatments with $n = 11$ or $n = 51$ observations per treatment. ANOVA results using the raw data indicated that the distribution of the empirical F ratio agreed quite well with the nominal distribution at significance levels of $\alpha = 0.05$ and $\alpha = 0.01$.

The results of the Hsu and Feldt study apply to the case where the general null hypothesis is true, *i.e.*, where treatment means are all equal. Therefore, the assumption of homogeneity of variance was not violated. The study of Lunney extended the results of Hsu and Feldt and also examined power of the F-test, the latter involving heterogeneity of cell variances since no variance stabilizing transformation was used. Lunney investigated values of $p = 0.1, 0.2, 0.3, 0.4$, and 0.5. Several 1, 2 and 3 dimensional table layouts were investigated with sample sizes per cell ranging for 3 to 31 in steps of 4. The results presented in Lunney's paper are averages of results from various layouts so interpretation and comparison to other studies is difficult. The main conclusion of the study was that the F-test is robust for binomial probabilities ranging from 0.2 to 0.8 given that cell sample size is fixed and under the condition that the number of degrees of freedom for the within cell variance is equal to 20 or more. Power results are not displayed by Lunney but it was commented that when the above conditions are satisfied the observed power is close to the nominal power.

Other studies of the robustness of ANOVA, but not for the binomial data situation, are reported in [RoK77] and [Nor52]. The Rogan study investigates the robustness of ANOVA to variance heterogeneity when sample sizes are equal and data is normal. The main result concerns a quantification of how much heterogeneity of variances is too much. The degree of heterogeneity is measured with the coefficient of variation of the variances, C_{σ}^{2}, the standard deviation of the variances divided by the mean of the variances (see [KeF71]). The results of the study suggest that ANOVA F is robust when $C_{\sigma}^{2} < 0.80$. The Norton study demonstrated similar robustness of ANOVA F in the case of heterogeneity of variances. Also, Norton showed that ANOVA F is robust to nonhomogeneous cell distribution shapes, regardless of whether heterogeneity of cell variances is present.

4.3. Comparisons of Loglinear Analysis and ANOVA

No extensive studies directly comparing ANOVA to loglinear analysis were found in the literature surveyed for this thesis. Cox [Cox70] briefly comments on the results of some empirical investigations comparing analyses of binary data using four different response curves. An example is presented that compares the stimulus-response curves for linear, logistic, angular, and standardized normal scales. His results suggest there is little difference in the methods. The logistic and normal curves show the closest agreement. The angular (arcsine transformation) and logistic curves are in close agreement for probabilities in the range 0.2 to 0.8 (however, scale constants were chosen for the example so that all curves would agree at the $p = 0.8$ point).

A single case example is detailed in [BFH75] where a set of data is analyzed using logit models and using ANOVA with the arcsine transformation of Freeman and Tukey. The conclusions about main effects and interactions are the same for both methods with the exception of a single interaction term being found just slightly significant by the ANOVA procedure but not included in the best fitting logit model. Though the authors conclude that the effect can be ignored in the ANOVA model, it is worth noting this incident of discrepancy between the two methods since other single case examples have demonstrated that ANOVA may lead to inclusion or rejection of interaction effects in opposition to results of the corresponding logit analysis.

Hsu and Feldt include a limited discussion in [HsF69] for some results comparing the analysis of 2-way tables of counts using the usual χ^2-test of independence and ANOVA. Several arguments are presented in favor of ANOVA for this situation. However, no mention is made using loglinear analysis (logit) for the data, which for the situation they examined would be more appropriate since the χ^2-test of independence is not for hypotheses concerning dependency as is the ANOVA F-test. Because this study was carried out before loglinear analysis had been broadly publicized or even fully developed, perhaps the authors were unaware of its applicability in place of the χ^2-test of independence. In a study by Knoke [KnB80] loglinear models are compared to dichotomous dependent variable multiple regression models using dummy variables. Two specific cross tabulations are analyzed using both methods. Knoke reiterates the observation of Goodman [Goo78] that the two methods can lead to different conclusions when the dependent dichotomy falls outside the range of 0.25 to 0.75. In the first example, one on voter willingness to legalize abortion, the response proportions are within the aforementioned range, and the two methods lead to the same conclusions. In the second example, voter willingness to legalize use of marijuana, several of the response proportions fall outside the range. For this case, the methods yield different conclusions about the data. The dummy variable regression identifies significant dependent variable interaction terms whereas the loglinear modeling does not. Knoke does not go into detail to offer an explanation for the differences other than noting the differences occur when proportions for the response variable are extreme and commenting the assumptions about normality of the regression error terms is violated.

In [Mag78] Magidson makes a similar comparison of dichotomous dependent variable regression analysis and logit modelling resembling that of Knoke. Again the situation where the dependent response proportions are extreme is identified as one where differing results may occur for the two methods. Magidson examines in detail data for predicting camp preference of American soldiers. In addition to this example, Magidson includes a separate discussion about the differences in the definitions of effects for both types of modelling. Clear examples of cases where the two methods lead to different conclusions concerning interaction effects in 2×2 tables are presented in the concluding remarks. These are worth reviewing and are adapted here to help emphasize and clarify the basic differences in what interactions are in the two types of models. In the example tables below success probabilities are in parentheses next to the corresponding odds of success to failure. What should be noted here is that there exist data where interaction effects are present in analysis using probabilities and absent in analysis using odds (and *viceversa*) even though both use the same data.

1.04 (0.51)	0.27 (0.21)
1.00 (0.50)	0.25 (0.20)

Table 4.1 Magidson example 1: row effects (very small for both probability and odds), column effects (small for both probability and odds), and interaction effects (none in probabilities, very small in odds).

1.04 (0.51)	0.011 (0.011)
1.00 (0.50)	0.001 (0.001)

Table 4.2 Magidson example 2: row effects (very small in probabilities, large in odds), column effects (large in probabilities, very large in odds), and interaction effects (none in probabilities, large in odds).

11.0 (0.917)	0.011 (0.011)
1.00 (0.500)	0.001 (0.001)

Table 4.3 Magidson example 3: row effects (small in probabilities, large in odds), column effects (very large in probabilities, extremely large in odds), and interaction effects (large in probabilities, none in odds).

9.00 (0.9)	X
1.25 (0.2)	4.0 (0.8)

Table 4.4 Magidson example 4: Every value of X in the additive probability model must show an interaction effect whereas the odds model shows no interaction effect for $X = 144$ (odds) or 0.993 (probability).

Chapter 5

Simulation Design

The Monte Carlo technique was chosen as the method by which to generate empirical results for comparing loglinear analysis to classical fixed-effects ANOVA when both are competing methods for testing the same hypotheses on an appropriate set of data. Monte Carlo studies have become a common way to empirically test the agreement of applied statistics with the nominal theory. This study was carried out to examine the levels of agreement over varying conditions and to then utilize the generated results to draw conclusions about which form of analysis exhibits superior performance over the conditions that were tested.

In this chapter the tests used to measure the goodness-of-fit for each method of the summary statistics to their corresponding theoretical values are first explained and justified. Next, the design of the Monte Carlo study is detailed. This includes rationale for the probability layouts in the tables, hypotheses tested, simulation size, and discussion of the conditions that were varied as they relate to both loglinear analysis and ANOVA.

5.1. Hypothesis Testing

In practice, the two most commonly used statistics for assessing the goodness-of-fit of a loglinear model are the Pearson X^2 statistic and the likelihood ratio statistic G^2. Both have asymptotic chi-squared distributions under the null hypothesis. For example, given a 2×2 table of counts that are drawn from a multinomial distribution with population probabilities $(p_{11}, p_{12}, p_{21}, p_{22})$ corresponding to the cells of the table, the completely null hypothesis that the probabilities are all equal is tested with X^2 and/or G^2. If the null hypothesis is true then these statistics are distributed approximately as a chi-squared distribution with three degrees[5] of freedom. The test for rejection of the null hypothesis may be made by selecting the desired level of significance, often denoted by α, and comparing the value of the calculated statistic with the critical value corresponding to the selected level of significance.

---

[5] The fully saturated log-linear model for an IxJ table is $log(m_{ij}) = u + u_{1(i)} + u_{2(j)} + u_{12(ij)}$. The constraints of the null hypothesis that $u_{1(i)} = u_{2(j)} = u_{12(ij)} = 0$ allow $(I-1)+(J-1)+(I-1)(J-1)$ parameters to vary. Hence, in this example there are $v = (2-1)+(2-1)+(2-1)(2-1) = 3$ degrees of freedom.

Similarly, ANOVA models are tested using F statistics that are calculated from the data. Evaluation of ANOVA models involves the calculation of several F statistics, one for each effect that is included in the null hypothesis, whereas the goodness-of-fit for loglinear modeling can be based upon a single summary statistic that takes in the entire model. However, thorough study using loglinear analysis will usually embody much more than the calculation of X^2 and G^2. Procedures for model selection, data rearrangement, and assessment of internal goodness-of-fit, to name a few, are a common part of loglinear analysis. For the most part, applied studies using either loglinear analysis or ANOVA are not confined to calculations of their basic summary statistics, X^2, G^2, and F-ratios, respectively. Yet, since they are the backbones of the methods under investigation, this study bases the comparison on results stemming from them alone.

For each table of counts generated in the simulation X^2, G^2, and the appropriate F statistics are computed. These statistics are then examined to see how well they approximate their corresponding asymptotic distributions. The methods used in this study to check the fit to the approximations are graphical analysis, the Kolmogorov-Smirnoff test, and calculation of Type I and Type II error rates (though power is reported in place of Type II error).

5.2. Kolomogorov-Smirnov Test

The Kolmogorov-Smirnoff test is used to test the null hypothesis of an unknown distribution being equivalent to a known distribution. The test is defined as

reject the null hypothesis if $\max|F(x)-\hat{F}(x)| > \delta$.

The left hand side of the inequality represents the maximum absolute deviation of the empirical cumulative distribution function $\hat{F}(x)$ from the theoretical cumulative distribution function $F(x)$ assumed by the null hypothesis. δ is the critical value corresponding to a selected level of significance. The test is applicable only for continuous underlying distributions. Finally, a non-significant statistic does not imply that the sample distribution is the null distribution; there could be several distributions that do not differ significantly from the null distribution.

The Kolmogorov-Smirnoff test was chosen to be included in this study since it affords a simple, comprehensive measure of goodness-of-fit of the empirical distributions of each of the summary statistics to their corresponding theoretical distributions. The Kolmogorov-Smirnoff statistic was found for each empirical distribution. Because it is a broad spectrum test, the results were used solely to detect cases of significant disagreement between empirical and theoretical results.

5.3. Type I Error Rates

A Type I error is defined as the case where the null hypothesis is rejected when in fact it is true. In the context of this study Type I error was examined via cumulative probability percentage points of the empirical distributions of the summary statistics. A statistic is said to be significant at the α percent level when it exceeds the value δ where δ denotes the smallest number that has $1-\alpha$ percent of the values in the asymptotic distribution of the statistic less than or equal to it. The value δ is commonly referred to as a critical value.

A comparison of ANOVA and loglinear analysis was made by contrasting the empirical Type I error rates of each method with respect to chosen nominal rates. The nominal rates chosen were 0.10, 0.05, 0.025, and 0.01. For each nominal rate critical values were calculated by evaluating the inverse of the cumulative probability function. These critical values were then used to locate the corresponding empirical Type I error rates. For example, a nominal rate of 0.05 for an F distribution with two degrees of freedom in the numerator and 27 degrees of freedom in the denominator corresponds to a critical value of 3.35, rounding to two decimal places. This critical value is then used to find the largest observed F statistic, say, F_{crit}, less than or equal to 3.35 from the list of 10,000 calculated F statistics in a given simulation. The number of F statistics in the list that are greater than F_{crit} divided by the size of the list is the empirical Type I error rate corresponding to the nominal rate of 0.05.

The study of empirical Type I error rates with respect to the asymptotic (nominal) Type I error rates provides more than just another measure of the goodness-of-fit of the empirical distributions to the theoretical distributions. The critical values corresponding to the nominal rates (significance levels) are used to determine when to reject the null hypothesis in ANOVA and loglinear modelling When the empirical rates exceed the nominal rates the test statistic leads to rejection of the null hypothesis too often. When the empirical rates are less than the nominal rates they are said to be conservative in that they do not lead to rejection often enough. Examining the levels of conservativeness of the related summary statistics is a way of gaging the relative reliability of using loglinear analysis or ANOVA for testing a hypothesized model over the same set of data. Although loglinear analysis and ANOVA utilize summary statistics with different asymptotic distributions, each method can be compared via empirical Type I error rates. The measure of the relative conservativeness or liberalness of the statistics used for significance testing is the basis of comparison.

5.4. Type II Error Rates (Power)

A Type II error rate is the probability of not rejecting the null hypothesis at a given level of significance when in fact it should be rejected, that is, when it is not true. Type II error rate is directly related to what is called the power of a test. Power is defined as one minus the Type II error rate, or, the probability that the test will lead to correct rejection of the null hypothesis. Analytical measures of power are somewhat unwieldy to apply in practice since, in general, test statistics have a different distribution for each model outside the model upon which the null hypothesis is based. The problem is that given a specific null hypothesis, violation of the null hypothesis means that the distribution of the test statistic could be any of many that are possible aside form the null distribution of the null model. Therefore, power is specific to a given null hypothesis and given violation of that hypothesis.

This study compares ANOVA and loglinear analysis through empirical measures of power based upon several different violations of a set of null hypothesis. For example, power is examined for the case of a 2×2 table with the null hypothesis that no row, column, or interaction effects are present. The way to obtain an empirical measure of power is to have the simulated data be generated such that a known column effect is present and then test the hypothesis that the effect is absent. The method which exhibits larger degree of power, or, equivalently, lower Type II error rate, is then judged as being superior to the other method. As determination of power is most easily done through large scale simulation, it is natural to include power results in this Monte Carlo study.

5.5. Compared Orderings of Sample Tables

Even though the statistics from the two methods under investigation have different asymptotic distributions they should create similar orderings of the sample tables in each experiment when they are performing adequately. The orderings are based upon the ranks of the summary statistics for the tables. Hence, if the 199th table in the 2×2 design for row and column independence yields a X^2 value of 2.51 which has rank 1281 out of the 10,000 X^2 statistics calculated in that experiment, table 199 is given rank 1281. For this experiment the table ranks based on X^2 would be compared to the table ranks based on the F statistics for testing interaction effects. Suppose table 199 generates a F statistic for interaction of 1.23 which ranks 1269 among the other F statistics for interaction. Then, in this example, the two rankings of table 199 happen to differ somewhat.

An advantage with this type of comparison is that it is distribution-free. Direct comparison of the ranks avoids complicating issues concerning the two different underlying asymptotic distributions thus allowing for a simple comparison of the two methods which gave rise to the ranks. There are many well documented distribution-free methods available for analyzing rank data. Most address the question of whether the ranks result from the same underlying distribution of values upon which the ranks are based. Since there is no question about the different identities of the underlying distributions in this study, such methods that test for distribution equality are not considered.

The point of interest in this case is to investigate whether the two methods order the tables the same way. This is slightly different from distribution related investigation in that for cases where the summary statistics are not distributed particularly close to their corresponding asymptotic distributions for both methods, extensive comparison of the deviations from asymptotic values becomes somewhat intractable due to the size of the simulation. On the other hand, by freeing the analysis from the context of the theoretical distributions, the comparison of the similarity or difference of the behavior of the two methods when conditions are extreme or debilitating becomes more tenable.

Given the situation where both methods are failing, it is still of interest to know if they are failing in a similar fashion. Light and Margolis [LiM71] use the method of calculating the rank correlation coefficient as part of an examination of two different methods. In their case, they compared the statistic from their CATANOVA method to the Pearson $\mathbf{X}^2$ statistic derived from classical chi-squared testing for independence in two dimensional tables. The rank correlation coefficient has been selected as the measure to apply in this study.

For a given experiment a set of table ranks is generated for each method. The rank correlation coefficient is then calculated for the two sets. It is defined as

$$r_{xy} = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2-1)} = \frac{\frac{1}{N}\sum_{i=1}^{N}(X_i-\bar{X})(Y_i-\bar{Y})}{S_x S_y} = \frac{cov(X,Y)}{S_x S_y},$$

where X_i and Y_i are the two ranks given to the ith table, d_i is the difference between X_i and Y_i, N the number of tables generated, S_x the standard deviation of the X_i, S_y the standard deviation of the Y_i, and $cov(X,Y)$ the covariance of the X_i and Y_i pairs.

The range of r_{xy} is from -1 to $+1$. Absolute values close to one indicate that the two methods behave very similar, ranking the tables such that the ranks are linearly related. Values close to zero indicate that overall the assignment of ranks to the tables is not discernibly similar for the two methods. It should be noted that such a measure of behavior says nothing about the competing methods' relative performance in terms of their respective diagnostic capabilities. It is merely a way of detecting another degree of similarity or difference in the methods. Measurements of diagnostic power are necessarily dependent upon the specified asymptotic distributions since the methods rely upon significance testing of statistics following assumed distributions.

Still, the information given by the rank correlation coefficient can be useful when looked at in conjunction with the measures of diagnostic power to help identify situations which influence such power. For example, there may occur two situations where ANOVA fails quite drastically and where loglinear analysis succeeds adequately while, in the first case, the rank correlation coefficient is quite close to one and, in the other case, it is close to zero. In such a scenario the rank correlation coefficient suggests that special attention be given to the second case above since for it both the performance and the behavior of the two methods are quite different.

5.6. Simulation Design

This section presents the details of the structuring of the Monte Carlo study used to compare and contrast ANOVA and loglinear analysis. First, the overall aim and scope of the design are briefly outlined. Following that more general overview, detailed descriptions of the various facets of the design, such as table probability structures and sample sizes, among others, are supplied. Motivations and justifications are included along with those descriptions of elements which involved fundamental design decisions. Occasionally the reader is referred to the earlier chapters for details and/or examples, but, for the most part, an attempt has been made to keep the section self-contained.

5.6.1. Aim and Scope

As described in Section 3.6 of Chapter 3, ANOVA and loglinear analysis are being compared in this study only for the analysis of contingency tables where the observed counts in each cell represent the number of successes for a fixed number of Bernoulli trials, the number of trials being the same for each cell in a given table. This study generates large-scale simulations of experiments yielding such tables by assigning theoretical binomial distributions[6] to each cell of a table structured to test a given null hypothesis.

---

[6] If the probability of success for a cell is set at p, and n observations are to be made for that cell, then the n observations are sampled from a binomial distribution with parameters p and n to generate the count for that cell.

For example, to analyze a method's performance in testing the null hypothesis of no row effects, no column effect, and no interaction effects in a 2×2 table, each of the four cells of the table are assigned identical binomial distributions from which to randomly generate the observed cell counts. 10,000 tables of counts following that probability structure are then generated for use in conducting the various analyses detailed in Section 5.7.

The conditions which were chosen to be controlled and varied in this study are the table dimensionalities, the theoretical models with their corresponding null hypotheses, the size parameters of the underlying binomial distributions (fixed across cells of a table for a given model), and the magnitudes of the probabilities used to test a given null model. This last condition is also a way of controlling the size of the observed counts when testing a given model since reducing the probabilities is a way of generating smaller counts while still maintaining a given theoretical structural relationship among cell counts in a table.

Different combinations of these conditions were selected to define a given simulation. The conditions stated above were chosen as the simulation parameters since each will commonly vary in applied practice. Hence, some emphasis was placed on keeping this study focussed on obtaining results that could be useful to the practitioner. Though contrived layouts of extremes, such as table structures with very small probabilities, say, less than 0.05, do have theoretical merit, they were not investigated in this study. Such cases most often demand specialized modifications of existing techniques that become difficult if not impossible to generalize.

The extent of the simulation was limited to testing the main features of each method. The number of advanced modifications and extensions of both loglinear analysis and ANOVA is becoming increasingly large. Techniques other than overall model identification were not studied. Relative effectiveness of procedures such as parameter estimation and residual analysis have been left outside the scope of this study. Although these are important areas of investigation, because this was an initial comparison of the two methods, such extensive comparison was considered premature and best put off pending analysis of the preliminary results.

Enough was included in this general look to make a fair comparison of the two methods. The aim was to produce sufficient results to make a recommendation as to which procedure is superior overall. In order to achieve this goal, simulations were designed to compare and contrast the known weaknesses and strengths of both. Cases where one method fails and the other is robust were generated, as well as cases where both are known to perform questionably. Of greatest interest are the latter and those cases where the theory and prior results suggests both are on equal footing. For these situations the empirical study is most informative since no prior results exist which directly compare ANOVA and loglinear analysis in detail.

Another aspect of this study that influenced the design of the simulation was a desire to generate results that could be used to also corroborate previous results concerning issues of loglinear analysis and ANOVA independent of each other. This was not an initially planned part of the study but it followed easily from the rest of the work, and though several related studies have been carried out, the work has not been exhaustive. The cases in point here are the comparison of X^2 to G^2 and effectiveness of the two suggested "corrective" transformations investigated in this study, the arcsine transformation for ANOVA and the addition of 0.5 to counts for loglinear analysis. Also, a comparison of the stepwise forward and stepwise backward model selection strategies for loglinear analysis was included.

5.6.2. Environment

The simulations were performed using a VAX 11/780 and an Amdahl 580/5860 running under UNIX and MTS (Michigan Terminal System) operating systems, respectively. APL library routines were called for generating the sample tables and for calculating the theoretical cumulative probabilities and inverse cumulative probabilities. All sample statistics were calculated using programs written in C under a UNIX environment. Graphs were produced with a Calcomp plotter utilizing the TAG software package.

The binomial sampling was simulated by first specifying the population probability, P, and the sample size, N. The APL random number generator was then used to generate N five digit random numbers drawn from a uniformly distributed population over the interval from zero to one. The sample count for a table cell following a binomial population with parameters P and N was then taken to be the tally of the random numbers less than or equal to P. This process was repeated 10,000 times for each cell of each table examined in this study. The simulation size of 10,000 for each table was chosen so that the probability would be 0.99 or greater that the observed mean value of a cell (taken over the 10,000 tables) would be within 0.1 standard deviation of the population mean, NP, of the cell.[7] Also, the simulation size was selected to correspond to the size used in several related studies [HsF69], [RoK77] [WCT86] [Lar78] to allow for more direct comparison between studies. Finally, all numerical computations were carried out to an accuracy of six decimal places or more.

5.6.3. Table Dimensionality

Probability structures are examined only for 2×2 and 2×2×2 tables of counts. Actually, as noted in Section 3.6 of Chapter 3, these tables correspond to 2×2×2 and 2×2×2×2 tables with one configuration fixed when studied using loglinear analysis. The cardinality of each dimension is limited to two since the counts result from a dichotomous response variable. Thus, the sampling is in effect product multinomial sampling, the individual multinomials being binomials. For such a sampling scheme each multinomial is independent of the others in the table. Therefore, no asymptotic differences should be induced in either of the two methods by increasing dimension cardinality other than possible smoothing of results and the expected improvements related to the overall increase in sample size. If the individual multinomials were not restricted to being binomials, dimension cardinality would make a difference due to the theoretical constraint that cell population probabilities in a given multinomial must sum to one. In that case, higher cardinalities imply low cell probabilities or highly skewed cell probabilities, which in turn has the effect of increasing the likelihood of small counts appearing in the table unless the overall sample size is made large. The type

---

[7] This is found through a version of Tchebycheff's inequality,

$$prob.(|X-\mu|<k)\geq 1-\frac{\sigma^2}{k^2},$$

where X is any random variable, k a real constant greater than zero, μ the population mean of X, and σ^2 the population variance of X. Choosing X = sample mean, $k = .1\sigma^2$, and setting the right hand of the inequality to .99 allows a way to solve for N, the number of simulations, since the variance of the sample mean is $\frac{\sigma^2}{N}$.

of sampling scheme which forms the basis of the comparison of loglinear analysis and ANOVA, Bernoulli trials, removes as an issue the effect of the cardinalities of the table dimensions.

On the other hand, the number of dimensions relates to two areas of investigation in this study. First and most significant is that with increased dimensionality a greater number of models become possible for explaining the observed data. Higher order variable interactions are possible along with various combinations of lower order effects and interactions. Most of the studies referenced in the course of this research have examined tables of two dimensions or less, especially those related to ANOVA. For the studies involving loglinear analysis, most have had three or two dimensions as the upper bound on tables investigated. One motivation for choosing 2×2 and 2×2×2 (implying 2×2×2 and 2×2×2×2 for loglinear analysis) tables is that results for lower dimension tables can be related to previous studies while the higher dimension results can provide an extension. Authors of previous results in the literature suggest that the results for tables of lower dimensions should extend to higher dimensions.

In the case of loglinear modelling the study of many of the models in three or more dimensions requires the use of an approximation algorithm to calculate cell estimates. This study selects several models that require approximation of the estimates. In these cases where direct estimates do not exist, a precise analytical approach to studying the distributions of the test statistics under various violations or transformations becomes difficult if not impossible, making the alternative of an empirical investigation all the more worthwhile.

There is, however, sometimes a problem with high dimensionality. In applied situations, interpretation of models with a large number of significant terms becomes increasingly difficult and muddled. In some cases the researcher combines or collapses categories to aid with interpretation. In other cases tables are partitioned into smaller tables and then scrutinized. These are more advance areas of investigation and were not specifically addressed in this study. Tables of dimensionality higher than three for ANOVA layouts of the probabilities were not considered (no higher than four for the corresponding loglinear analysis).

5.6.4. Cell Sample Size

Since fixed effects ANOVA is being compared to loglinear analysis the the number of observations for each cell of a table is constant. Fixed effects ANOVA methods do exist for non-orthogonal designs [Coh82], but they are not considered in this study. To test the small sample and the large sample behavior, cell sample sizes of 10 and 40 are employed, respectively. These sizes are used for both dimensions of tables studied. The size of 10 observations per cell is considered to be about the smallest size of interest in this study for two reasons. First, the context of the comparison implies that for the loglinear analysis sampling size is fixed for cells[8]. This is due to the product-multinomial sampling scheme such that each table is stratified into binomials. As noted in 4.2 of Chapter 4 it has already been determined that both $\mathbf{X}^2$ and $\mathbf{G}^2$ are suspect when observed cell counts are 1 or 0. The smallest cell probability used in this study is 0.05. Hence, a cell sample size of 10 leads to an expected count of 0.5 in such a cell. Therefore, the small sample behavior of loglinear analysis is adequately examined by setting the number of observations per cell at 10 and choosing appropriately small probabilities.

The second reason for choosing the size of 10 stems from the results of given by Lunney [Lun70]. That study concluded that for ANOVA with a dichotomous dependent variable the number of degrees of freedom for within cell variance should not be less than 20 when the response probabilities for success are greater than 0.2. For cases where there are probabilities less than 0.2 it was suggested that 40 degrees of freedom are required to obtain reliable results. The size of 10 observations per cell corresponds to 40 degrees of freedom for within cell variance for 2×2 tables. Therefore, 10 observations per cell should be extreme enough to simulate known conditions where the tests used in both ANOVA and loglinear analysis exhibit less than adequate performance. Again, based on earlier findings, the cell sample size of 40 is about the point where significant changes in performance cease to occur. Further increases in sample size result in on slight improvements in the performances of the two methods.

---

[8] Actually it is fixed for pairs of cells in the loglinear formulation, but the information in one cell of the pair determines the value of the other cell.

5.6.5. Data Transformations

At several points in the history of the development of ANOVA and loglinear analysis, transformations of the raw data have been suggested to either make the data conform to underlying assumptions or to make adjustments to correct for violations. Two transformations are examined in this study: the arcsine transformation and the addition of 0.5 to cell counts. These are used with ANOVA and loglinear analysis, respectively. For each set of raw counts the transformations are carried out and the two analyses are executed. Results using the transformed data are then compared with each other and also to the results using the raw data.

5.6.5.1. Arcsine Transformation

The purpose behind the use of the arcsine transformation is to stabilize heterogeneous variances among cells of a table. The arcsine transformation is used for cases where the the heterogeneity results from data being sampled from differing binomial distributions[9]. Several different forms of the transformation have been proposed in the literature [BHH78], [Coc40], [Bar47], [Fis44], [FrT50]. The raw data is recorded as proportions so the scale is 0.0 to 1.0. For each of these transformations the variate on the new scale tends to a normal distribution as $n \to \infty$, where n is the number of trials. The transformations all generally succeed in stabilizing the variance, exhibiting inadequacies only with the tail area probabilities. When the sample sizes are small, the range of proportions over which the stabilization will result varies between the transformations that were reviewed.

The transformation chosen for this study is the one proposed by Freeman and Tukey,

$$\theta = \frac{1}{2}\left[\sin^{-1}\sqrt{\frac{x}{n+1}} + \sin^{-1}\sqrt{\frac{x+1}{n+1}}\right]$$

where x is the number of successes and n the number of trials. θ is measured in degrees in this study.

---

[9] The variance of a binomial distribution is a direct function of the mean: $\mu = NP$, $\sigma^2 = NP(1-P)$. Therefore, cells that have different population probabilities, P, will have different variances in the simulation.

Mosteller and Youtz [MoY61] show that with this transformation, for $n=10$ in the interval $0.1 \leq (prob\_success) \leq 0.9$, the ratio of the variance of θ to its asymptotic variance is nearly constant. For $n=50$ the ratio is constant in the interval $0.07 \leq (prob\_success) \leq 0.93$, the variance of θ being within 2% of its asymptotic value. They also show that the Freeman-Tukey transformation stabilizes the variance over larger intervals for small to moderate sized n than does the often cited transformation used by Bartlett [Bar47],

$$\theta = \begin{cases} \sin^{-1}\sqrt{\frac{x}{n}} & (1 \leq x \leq n-1) \\ \sin^{-1}\sqrt{\frac{1}{4n}} & (x = 0) \\ 90^\circ - \sin^{-1}\sqrt{\frac{1}{4n}} & (x = n) \end{cases}$$

5.6.5.2. Adding 0.5 to All Elementary Cell Totals

A difficulty that can arise in loglinear analysis is the occurrence of many zero counts in cells. Zeros in cells have the effect of distorting the X^2 and G^2 statistics (see Chapter 3.4 for details). Also, when the zeros occur such that a marginal total is also zero, the degrees of freedom need to be reduced, minus one for each zero marginal total of a single variable as that level is basically collapsed out of the table [Fie77].

Several places in the literature the adjustment of adding 0.5 to each cell has been suggested as a possible way to mitigate the problems caused by excessive numbers of observed zero counts. Goodman [Goo70] was the first to make the recommendation. Subsequent authors have since made reference to the practice. Also, this technique has been included as an option in some of the widely available loglinear computer software packages for contingency table analysis such as BMDP/4 and ECTA.

Theoretical justification of the adjustment is not given. Direct calculation of the X^2 and G^2 statistics on such modified data leads to statistics that are no longer distributed with the regular asymptotic chi-square distributions. Hosmane [Hos86] develops the corrected X^2 and G^2 statistics for the case of 2×2 tables. In a Monte Carlo experiment Hosmane shows that the unaltered X^2 for the raw data is in most cases superior to the modified G^2 and slightly better on average than the modified X^2. He suspects that the results generalize to higher dimension tables but at the same time feels further investigation is needed before sound conclusions can be made.

This study does not attempt to extend the investigation along the direction of Hosmane's study. To do that the modified statistics would need to be derived for each model studied. Instead, the regular X^2 and G^2 statistics are examined using the adjusted data. The intent is to note the effect of the adjustment as it is offered to applied practitioners in the existing software packages. Whether the addition of 0.5 to cell totals is beneficial is not obviously clear in light of Hosmane's results . This study aims in part then to clarify just what effect adding 0.5 has on the analysis.

5.7. Models and Probability Structures

This section presents the table probability structures used to generate the cell counts. In choosing the structures, first,the linear models corresponding to the null hypotheses under examination were selected. Probabilities were assigned to the cells of a table such that they agreed with the related linear model. Then, pseudo-random counts were generated based on those probabilities. The equations for the models are given in both ANOVA and loglinear formulations. The ANOVA equations are identical in form to the logit models which are, in effect, being fitted via the corresponding loglinear models. The probability structures are presented in tabular layouts. In what follows the cross between tables is used to indicate that the two tables are joined (are layers of the same table).

5.7.1. Completely Null Model

The completely null model applies when no effects of any kind are present. This model is used to check the Type I error levels. Table 5.1 shows the probability structures examined for two dimensional tables. The equations of the null model are

$$Y_{ijk} = \mu + \varepsilon_{ijk},$$

for ANOVA in two dimensions, and,

$$log\,(m_{ijk}) = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)}$$

for loglinear representation in three dimensions.

.05	.05
.05	.05

case 1.1

.1	.1
.1	.1

case 1.2

.175	.175
.175	.175

case 1.3

.25	.25
.25	.25

case 1.4

.375	.375
.375	.375

case 1.5

.5	.5
.5	.5

case 1.6

Table 5.1 Null model cell probabilities for 2×2 tables.

Figure 5.3 shows the probability structures for the three dimensional tables. The probabilities used are the same as those chosen for two dimensions. The corresponding equations for this higher dimensionality are

$$Y_{ijkl} = \mu + \varepsilon_{ijkl}$$

and,

$$log\,(m_{ijkl}) = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{4(l)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)} + u_{123(ijk)}\,.$$

.1	.1
.1	.1

X

.1	.1
.1	.1

case 2.1

.25	.25
.25	.25

X

.25	.25
.25	.25

case 2.2

.5	.5
.5	.5

X

.5	.5
.5	.5

case 2.3

Table 5.2 Null model cell probabilities for 2×2×2 tables.

5.7.2. Main Effects for One Variable

Tables that match the linear model having one main effect present and all other main effects and interactions set to zero provide the simplest setting in which to test Type II error levels (or power). Here there is only one violation of the complete null model to consider. For ANOVA the focus of attention is directed to the F-ratio that tests for the main effect that has been added. Any rejection of the complete null model stemming from other F-ratios would be spurious due to the source of rejection, even though rejection is the correct choice. For loglinear analysis, using stepwise forward model selection, the completely null model is compared to the model equaling the null model plus the term for the added main effect. Unlike the ANOVA case, significance testing of components in loglinear analysis involves fitting entire models and then checking the significance of differences. The G^2 statistic[10] of the completely null model is subtracted from the G^2 statistic of the model for one main effect. A nonsignificant difference results in a Type II error.

---

[10] The X^2 statistic cannot be used since it does not have the same additive properties.

.4	.4
.6	.6

case 3.1

.7	.7
.1	.1

case 3.2

.05	.05
.25	.25

case 3.3

.08	.08
.12	.12

case 3.4

.1	.1
.9	.9

case 3.5

Table 5.3 Cell probabilities of 2×2 tables for row main effects.

Stepwise backward model selection strategies are also examined. The models used to calculate the conditional G^2 are given in the next chapter.

The main effect for the first dimension was chosen as the one to be included. The dimension which contains the main effect makes no difference; the choice of first dimension is arbitrary. The same effect was chosen for both the 2×2 and 2×2×2 tables. The two and three dimensional models for one main effect in ANOVA format are

$$Y_{ijk} = \mu + \alpha_i + \varepsilon_{ijk}$$
$$Y_{ijkl} = \mu + \alpha_i + \varepsilon_{ijkl}$$

and in loglinear format are

$$log\,(m_{ijk}) = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)}$$
$$log\,(m_{ijkl}) = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{4(l)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)} + u_{14(il)} + u_{123(ijk)}$$

Tables 5.3 and 5.4 give the selected theoretical cell probabilities. Cases 3.1 and 4.1 have midrange probabilities which have been shown to allow high performance for both ANOVA and loglinear analysis. In the cases 3.5 and 4.3 the probabilities are at the extremes where both methodologies break down. Cases 3.3

and 3.4 are even more extreme in the probabilities and case 4.2 is mixed. Since the probabilities are not constant across all cells for each of the tables, the variances differ between cells in cases 3.2, 3.3, 3.4, and 4.2. Due to symmetry cases 3.1, 3.5, and 4.3 allow for row effects to be present without and heterogeneity of cell variances.

.4	.4
.6	.6

X

.4	.4
.6	.6

case 4.1

.7	.7
.1	.1

X

.7	.7
.1	.1

case 4.2

.1	.1
.9	.9

X

.1	.1
.9	.9

case 4.3

Table 5.4 Cell probabilities of $2\times2\times2$ tables for row main effects.

5.7.3. Full Main Effects Model

The inclusion of all main effects without interaction effects is investigated with these models. No interaction in the additive probability model (ANOVA) does not necessarily mean that log-odds interaction effects are absent. For these models log-odds interaction effects are present in cases 5.2, 5.3, 6.1, and 6.2. The other cases are of interest also since the loglinear models used to fit the logit models were chosen so that any relationships between the dependent variables is taken out when calculating the conditional G^2 statistics. This modellingtechnique is used regardless of the true presence or absence of relationships between dependent variables). Whether this automatic "partialing" has any effect on the Type I error or power might show with these other cases.

.3	.5
.5	.7

case 5.1

.05	.25
.25	.45

case 5.2

.06	.1
.1	.14

case 5.3

.1	.7
.3	.9

case 5.4

Table 5.5 Cell probabilities for 2×2 tables with both row and column effects present.

For the three dimensional case, direct maximum likelihood estimates are not possible when fitting the corresponding loglinear model. Hence, the effects, if any, of using approximate maximum likelihood estimates found through application of the Stephen-Deming [DeS40] iterative proportional fitting algorithm can be noted.

The equations for the mutual independence models are

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$$
$$Y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + \varepsilon_{ijkl}$$

for two and three dimensions with ANOVA, and,

$$log\,(m_{ijk}) = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)}$$
$$log\,(m_{ijkl}) = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{4(l)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)} + u_{14(il)} + u_{24(jl)} + u_{34(kl)} + u_{123(ijk)}$$

for loglinear representation of the corresponding logit models. The simulation probability structures for the tables are displayed in Table 5.5 and Table 5.6.

.2	.4
.4	.6

X

.4	.6
.6	.8

case 6.1

.1	.5
.3	.7

X

.3	.7
.5	.9

case 6.2

Table 5.6 Cell probabilities for 2×2×2 tables with both row, column, and layer effects present.

Chapter 6

Simulation Results

The results of the Monte Carlo study are presented in this chapter. Relative performances of the test statistics for ANOVA and loglinear analysis are compared and contrasted with respect to the simulation design detailed in Chapter 5. Most of the results are in terms of Type I error levels and power. The following comparisons are made:

- ANOVA F using raw data versus ANOVA F using the same data following application of the arcsine transformation of Tukey and Freeman;
- X^2 and G^2 of loglinear modelling using raw data versus the same statistics after 0.5 has been added to each cell total [Goo70];
- X^2 versus G^2 as measures of overall goodness-of-fit for loglinear modelling
- conditional G^2 of loglinear analysis versus ANOVA F (testing the same null hypotheses). Included here also is the additional comparison of forward versus backward stepwise selection strategies for loglinear modelling.

For the most part, performance is assessed in terms of how close the statistics match the upper 10% of their respective cumulative reference (nominal) distributions. The study of power was confined to the case where $\alpha = 5\%$ to keep the size of the study to a manageable level. Though the bulk of the results pertain to the upper percentage point behavior of the statistics, several complete distribution results are given in the form of graphs and measures of correlation.

6.1. Interpretation of the Graphs

Graphs of the nominal versus cumulative distributions of the statistics were produced for most of the simulation experiments. Most turned out to be repetitive, though in several cases interesting results did surface. Figure 6.1 depicts a detailed example of the type of graphs referred to in this section. The plots were made of the percentage points of the empirical cumulative distribution versus the percentage points of the reference cumulative distribution for the various summary statistics. In the simulation 10,000 tables of randomly generated samples were obtained for each test case. Sample statistics were calculated for each table of sample counts. Each collection of 10,000 statistics was then sorted in non-decreasing order. Each statistic was then used to create a point in the plot of a graph such as that found in Figure 6.1.

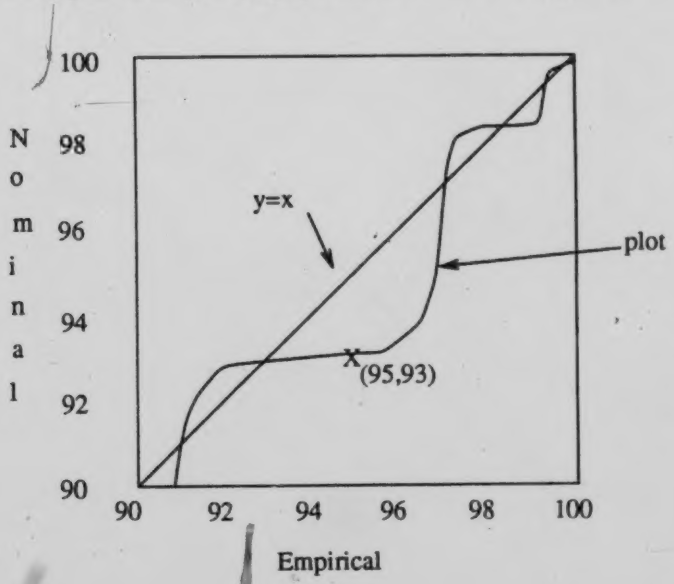

Figure 6.1 Example plot of nominal versus empirical cumulative

Assume the above plot is for the likelihood ratio statistic G^2. For the point X in Figure 6.1, the abscissa value, 95, represents the proportion of the 10,000 observed G^2 statistics less than or equal to the G^2 statistic in the simulation, say, G^2 with a value of 4.5. The ordinate value, 93, represents the probability of a value in the reference distribution being less than or equal to the value of the G^2 statistic. The ordinate (in this case 93) for G^2 equal to 4.5 is found by evaluating

$$\int_0^{4.5} f(x)\,dx \,,$$

where $f(x)$ is the probability density function of the reference distribution of the statistic being studied (either χ^2 or F in this study).

Visual inspection of the resulting plot reveals how well the empirical cumulative distribution agrees with the reference cumulative distribution. A straight line with slope equal to 1 indicates perfect agreement. Deviations of the plot above the line $y = x$ indicate that the percentage points of the reference distribution exceed those of the empirical distribution, while deviations in the opposite direction indicate that the reverse is true. The interpretation of point X is that the 5% critical value of the empirical distribution is in fact the 7% critical value of the reference distribution. Hence, at this point the test statistic is conservative in that a test of significance with it using tables of the nominal distribution will lead to rejection of the null hypothesis 2% less often than is nominally expected.

Ideally, the empirical distribution should match the nominal distribution. Emphasis of the Monte Carlo study was placed upon examining situations where such a match was not highly likely so that deviations could be compared. Plots deviating with regularity to both sides of the line $y = x$ are indicative of discreteness for the empirical distribution. Plots that are consistently above or below the line mean the statistic leads to a liberal or conservative test, respectively. Graphs of the entire cumulative distributions as well as graphs of just the upper 90 percentiles were plotted, the latter emphasizing the region where most significant testing is based.

6.2. Tabled Results

Results were collected in terms of actual significance levels using critical values of nominal levels for $\alpha = 10\%, 5\%, 2.5\%, 1\%$. Binomial standard deviations for realized significance levels of magnitudes 1%, 2.5%, 5%, 10%, 20%, and 50% are .10%, .16%, .22%, .30%, .40%, .50%, respectively.

Many of the results presented in this section are in terms of the number of standard deviations of the actual levels from the nominal levels. The reason for this transformation is to make the results more directly comparable. Since four different significance levels are being compared side by side, transformation of the deviations from nominal levels onto the same scale allows for fair comparison (in terms of magnitude) between different significance levels and thus a clearer view of the upper percentage point behavior. Furthermore, the results in this form are z-statistics and can be referred to normal distribution tables for tests of significance. Throughout this chapter a result is "significant" when its absolute value is 2 or more. This is approximately the critical value for the 5% level of significance for the z-distribution. The sign indicates direction of deviation. A negative sign means the observed value is below the nominal value (*i.e.*, test is conservative). Positive entries (unsigned) mean the opposite is true (liberal test).

Results are presented by statistic and case. Table 6.2 gives a quick reference guide of the cases in terms of the underlying probability structures of the test tables. Cases are referred to by number throughout this chapter.

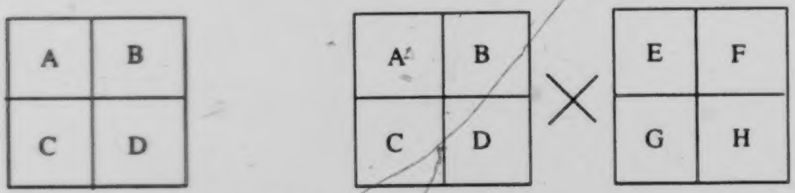

Table 6.1 2 × 2 table probabilities (A, B, C, D) and 2 × 2 × 2 table probabilities (A, B, C, D, E, F, G, H)

dimensions	effects present	case	cell probabilities
2×2	none	1.1	(.05, .05, .05, .05)
		1.2	(.1, .1, .1, .1)
		1.3	(.175, .175, .175, .175)
		1.4	(.25, .25, .25, .25)
		1.5	(.375, .375, .375, .375)
		1.6	(.5, .5, .5, .5)
2×2×2	none	2.1	(.1, .1, .1, .1, .1, .1, .1, .1)
		2.2	(.25, .25, .25, .25, .25, .25, .25, .25)
		2.3	(.5, .5, .5, .5, .5, .5, .5, .5)
2×2	row	3.1	(.4, .4, .6, .6)
		3.2	(.7, .7, .1, .1)
		3.3	(.05, .05, .25, .25)
		3.4	(.08, .08, .12, .12)
		3.5	(.1, .1, .9, .9)
2×2×2	row	4.1	(.4, .4, .6, .6, .4, .4, .6, .6)
		4.2	(.7, .7, .1, .1, .7, .7, .1, .1)
		4.3	(.1, .1, .9, .9, .1, .1, .9, .9)
2×2	row and column	5.1	(.3, .5, .5, .7)
		5.2*	(.05, .25, .25, .45)
		5.3*	(.06, .1, .1, .14)
		5.4	(.1, .7, .3, .9)
2×2×2	row col layer	6.1*	(.2, .4, .4, .6, .4, .6, .6, .8)
		6.2*	(.1, .5, .3, .7, .3, .7, .5, .9)

Table 6.2 Summary of probabilities used in test cases. Effects are noted in terms of additive probability model. Cases marked with * have large log-odds interaction effects. In all other cases the presence of log-odds effects corresponds with the labels in column two.

6.3. Arcsine Transformation

The arcsine transformation is used with ANOVA when the original scores are believed to be drawn from binomial populations. Application of the transformation results in new scores which have near constant variance (for fixed sample size, n) and are approximately normally distributed; that is, the problems of heterogeneity of variance and non-normality of cell populations are remedied. In experiments where the cell populations (binomial) do in fact have equal probability parameters (equal in the sense that the smaller of the binomial parameters, p and q, for each cell population is the same for each cell), heterogeneity of variance should, in theory, not exist, and the expected benefit of applying the transformation would be the improved normality of the cell populations.

This section details the results of the study aimed at assessing the effectiveness of the arcsine transformation in the aforementioned situation. Results comparing the empirical and nominal distributions of the F statistic are presented for the cases where the null hypothesis in the F-test is true by design of the Monte Carlo simulation. Discussion of the outcomes for the cases where the null hypothesis is designed to be false (power study) follows in Section 6.3.2.

6.3.1. Type I Error – Upper Percentage Points

Table 6.3 displays the results for 2 dimensional cases where no treatment effects are present. This table gives results of the F-test for row effects grouped together so that the effects of success probability size and sample size can be seen more easily. The case results for column and interaction effects are similar and can be found in the tables comparing ANOVA and loglinear results (Tables 6.15 to 6.26).

Increased cell sample size leads to actual levels that are close to nominal levels. Without the arcsine transformation 11 out of 24 levels are greater than 2 standard deviations from nominal for $n = 10$ observations per cell, while only 4 are greater than 2 standard deviations for $n = 40$. When the transformation is applied to the raw scores 17 levels deviate from nominal by more than 2 standard deviations for $n = 10$. For $n = 40$ there are 4 instances where the levels are significantly different from the nominal levels.

STANDARD DEVIATIONS OF ANOVA F FROM NOMINAL SIGNIFICANCE LEVELS

			nominal level (α)			
n (cell)	case	prob	0.10	0.05	0.025	0.01
10	1.1	.05	–14.9	–17.7	–14.5	–9.9
	1.2	.1	2.2	–4.7	–7.4	–7.2
	1.3	.175	1.0	0.6	–1.7	–1.4
	1.4	.25	0.3	0.5	0.9	1.3
	1.5	.375	–3.7	–0.8	1.0	0.4
	1.6	.5	–6.2	–2.8	–1.5	0.0
10†	1.1	.05	–17.8	–18.1	–15.0	–9.9
	1.2	.1	–0.7	–5.7	–8.5	–7.3
	1.3	.175	4.1	2.1	–1.5	–1.2
	1.4	.25	4.0	4.0	2.4	2.9
	1.5	.375	–4.2	–1.0	1.0	0.0
	1.6	.5	–6.8	–3.7	–3.2	–1.5
40	1.1	.05	5.1	–3.3	–3.4	–3.3
	1.2	.1	–0.5	–0.5	–1.9	–1.4
	1.3	.175	0.2	–0.3	–0.2	0.3
	1.4	.25	1.0	0.6	–0.2	–0.1
	1.5	.375	–1.2	–0.3	1.9	0.7
	1.6	.5	–0.7	–1.2	–1.5	–0.1
40†	1.1	.05	7.4	2.6	1.7	0.8
	1.2	.1	–0.2	–0.5	0.3	0.8
	1.3	.175	–1.1	–0.6	–0.8	–0.1
	1.4	.25	–0.5	–0.3	–0.3	–0.2
	1.5	.375	–2.0	–1.3	0.6	–0.4
	1.6	.5	–0.6	–1.2	–2.1	–0.2

† Entries where arcsine transformation was carried out.

Table 6.3 Standard deviations from nominal Type I error rates for ANOVA test of null hypothesis of no row effects

Upon closer inspection the results reveal that at $n = 10$ the pattern of deviations is similar between the tests using the transformed and raw scores. In the vicinity of the extremes and middle of the cell success probabilities ($p = 0.05$, $p = 0.5$) the realized levels are more conservative than the nominal levels. Agreement with nominal levels is closest when cell probabilities are in the region $0.175 \le p \le 0.375$ and when the arcsine transformation is not applied. Only 1 of the 12 empirical levels investigated in this region is greater than 2 standard deviations from the nominal level (roughly, significant at the 5% level).

The region $0.175 \le p \le 0.375$ also contains the smallest deviations for the transformed scores, but 7 empirical levels are clearly significant (at the 5% level). Also, the deviations are at a maximum within this region at $p = 0.25$ for the transformed scores, whereas, for the raw scores, $p = 0.25$ is the probability that has realized levels that are closest to nominal levels. Thus, at $n = 10$ superior results are observed for non-transformed scores, though the pattern of deviation in magnitude and direction is similar for both types of scores.

No such pattern is evident in the results when $n = 40$. Aside from the case where $p = 0.05$, the empirical levels are quite close to nominal levels and the discrepancies are most likely attributable to the inherent randomness of the study. When $p = 0.05$ the transformation leads to consistently more liberal levels than does the use of the raw data, and compared to nominal levels the transformation leads to liberal levels while the raw scores lead to conservative levels (except for the nominal 10% level).

Examination of the cumulative plots (Figure 6.2) shows that the arcsine transformation results in a more smoothed (but only slightly) cumulative distribution of F-ratios when the number of observations per cell is 40. Still, the cell probability of 0.5 leads to a rather discrete plot. This is very odd since one would expect such a population to be closer to normal than the other binomial populations since it is symmetric. Even more striking is that when the probability is 0.1 the graph (Figure 6.3) is nearly a straight line, what one would have expected to occur with $p = 0.5$ before occurring with $p = 0.1$. For probabilities less than 0.1 (graphs not shown) things do indeed begin to break down as should be expected, and both transformed and raw scores lead to more discrete graphs and larger discrepancies from the nominal cumulative distribution.

At 10 observations per cell the transformation results in a cumulative distribution of F-ratios that is more discrete than it is without the transformation; the steps become clearly defined. At $p = 0.1$ the F cumulative for the raw score ANOVA has steps, but they are more rounded than is the case when the arcsine transformation is used (Figures 6.4, 6.5). When $p = 0.25$ both transformed and raw scores lead to graphs (Figures 6.6, 6.7) that are relatively smoother (compared to $p = 0.1$ and $p=0.5$) with smaller steps. Still, the arcsine transformation brings out clearer steps. At $p = 0.5$ the raw scores ANOVA graph (Figure 6.8) is clearly discrete, much more than those for $p = 0.1$ and $p=0.25$. With use of the arcsine transformation the graph (Figure 6.9) is even more discrete in that the corners of the steps are not as rounded, instead having one or two short steps between the large ones. All the above results hold in both the 2 and 3 dimensional layouts.

It is apparent from these results that, for small samples, the arcsine transformation results in a cumulative distribution that is more discrete than the distribution obtained by using the raw binomial scores. A possible explanation for this is that the transformation maps the raw scores into a bounded interval thereby making the results more susceptible to discreteness due to rounding errors. Of course, as the asymptotic theory shows, this discreteness should vanish when the sample size is large.

Still, this does not explain the increased discreteness observed with use of the raw scores as binomial probabilities approach 0.5. This is baffling to the author, the only suspicions being that it may in some way be related to the increased symmetry of the distribution from which the scores are drawn (which does not seem a likely cause), or to an unexpected property of the sampling model (like the unexpected, but theoretically backed result in [Fel57] where lengths of betting leads in binomial experiments are shown to be longest when when $p = 0.5$), or to an artifact in the pseudo-random number generation. (The first four moments about the mean were calculated for each sample and did not show any great disagreement with theoretical moments. The third and fourth moments about the mean showed discrepancies in some cases, but always appeared to be approaching those of the normal distribution).

Results from the analysis of $2 \times 2 \times 2$ tables are similar to those for 2×2 tables (see Tables A2.1, A2.2, and A2.3 in Appendix 2). The main difference is that, overall, the empirical levels are closer to the nominal levels than they are for the 2×2 design. This agrees with the observation that the tests improve with increased sample size (and larger degrees of freedom [Lun70]).

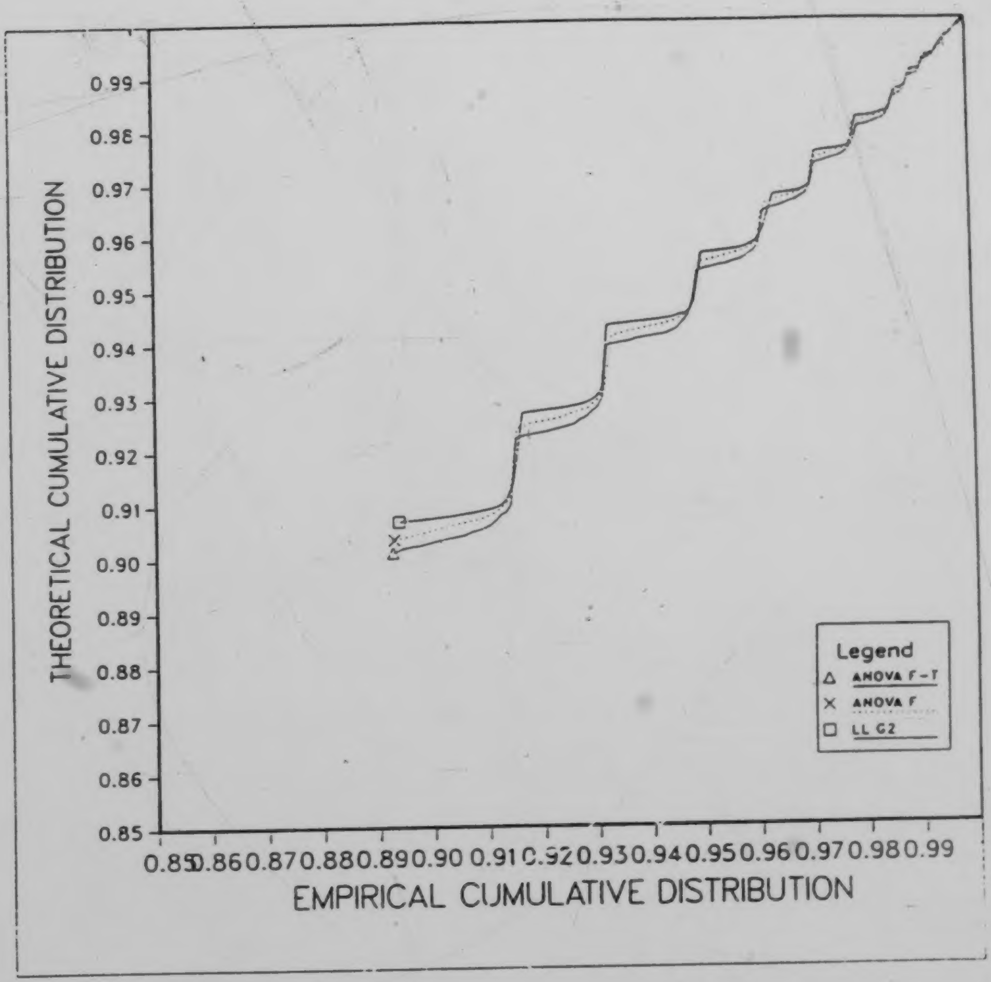

Figure 6.2 ANOVA F with arcsine transformation; ANOVA F without arcsine transformation; conditional G^2 backwards: testing for row effects, 40 observations per cell, 2×2×2 table, $p = 0.5$ (case 2.3).

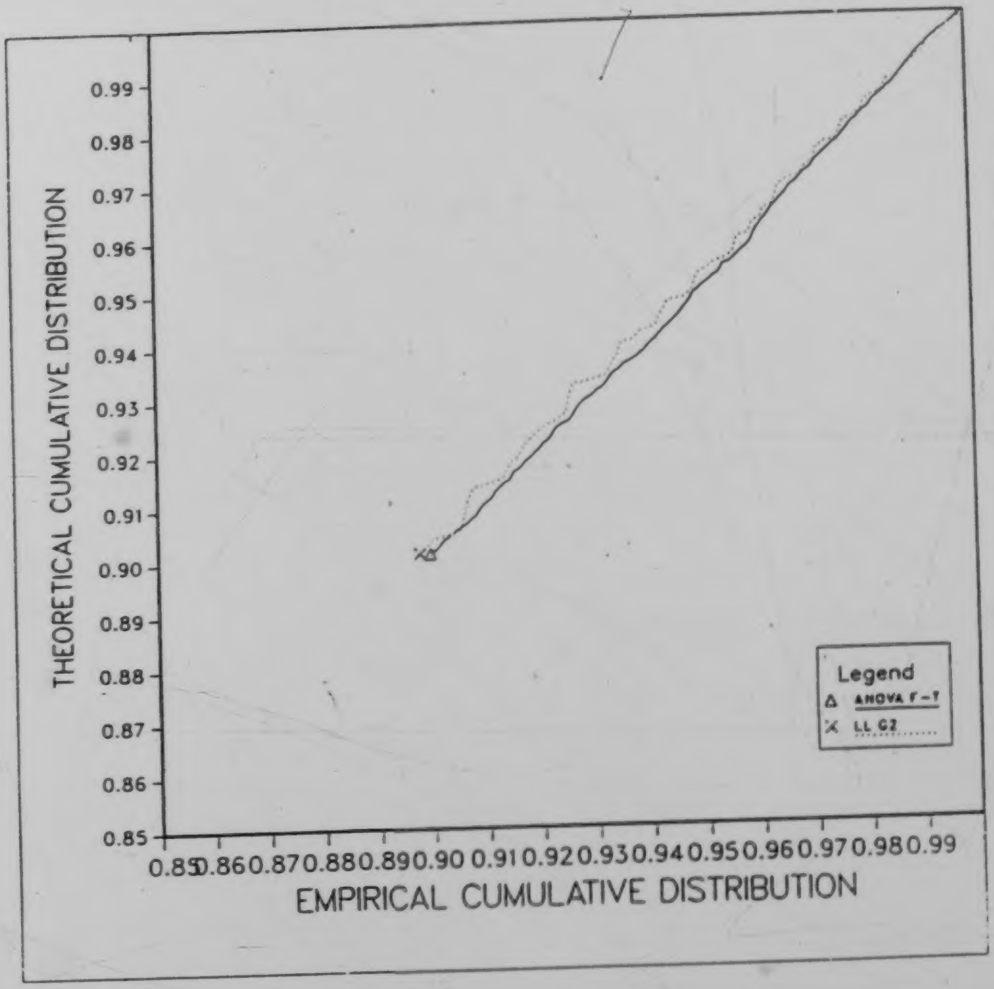

Figure 6.3 Conditional G^2 backward and ANOVA F with arcsine transformation testing for row effects: 2×2×2 table, 40 observations per cell, $p = 0.1$ (case 2.1).

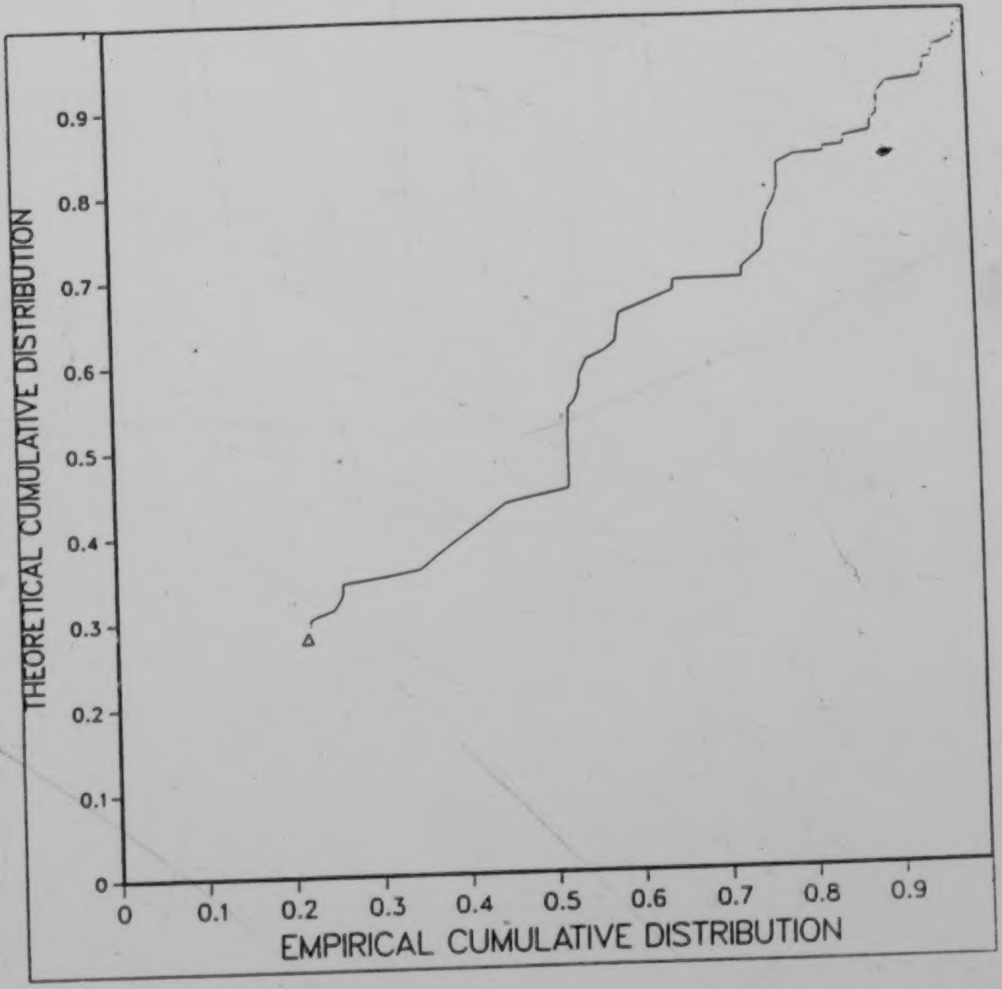

Figure 6.4 ANOVA F testing for column effects with no effects present: $p = 0.1$, 10 observations per cell, 2×2 table (case 1.2).

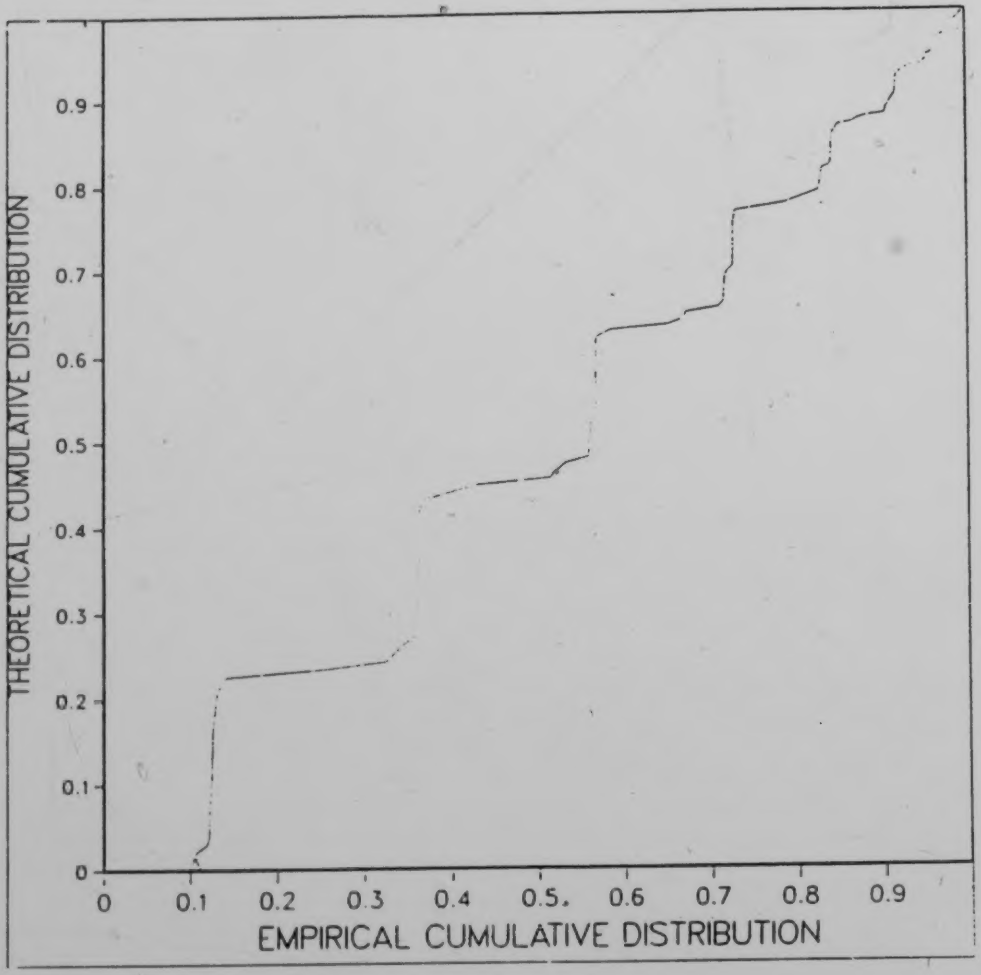

Figure 6.5 ANOVA F with arcsine transformation testing for column effects with no effects present: $p = 0.1$, 10 observations per cell, 2×2 table (case 1.2).

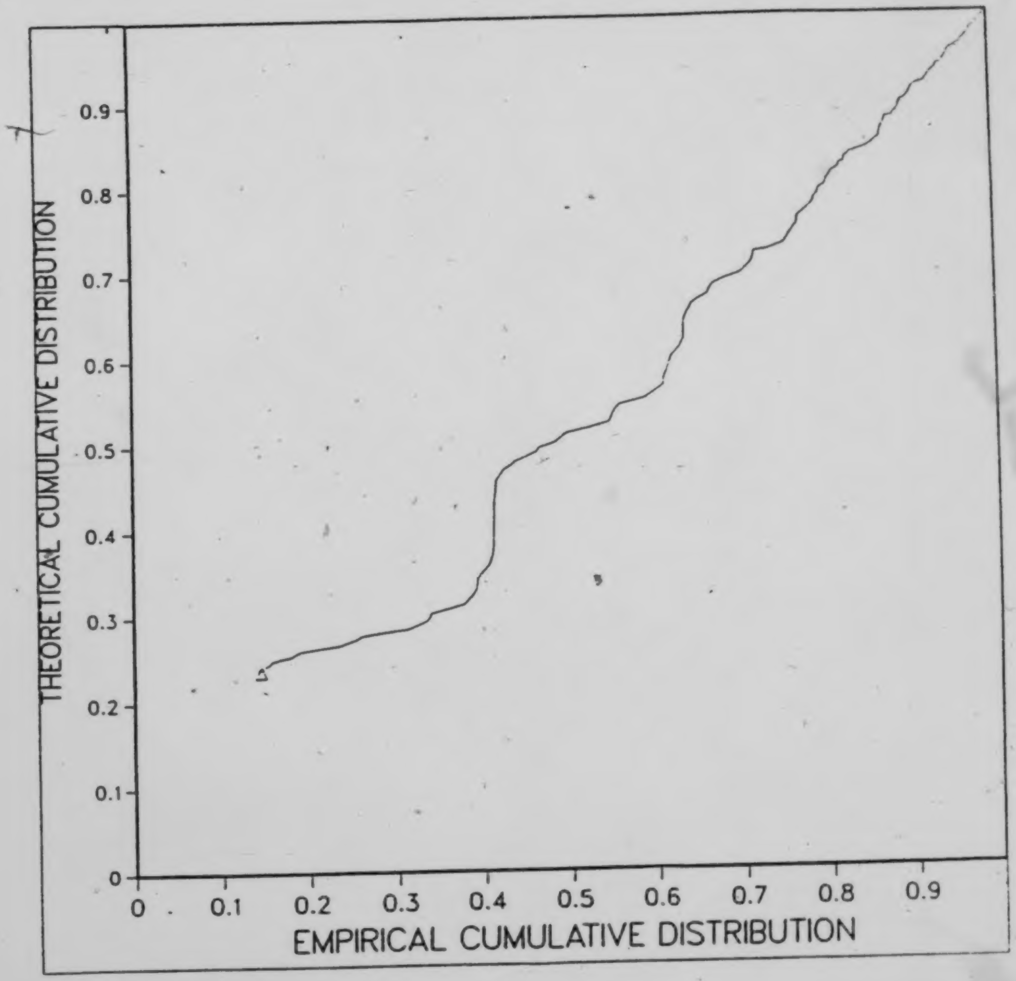

Figure 6.6 ANOVA F testing for column effects with no effects present: $p = 0.25$, 10 observations per cell, 2×2 table (case 1.4).

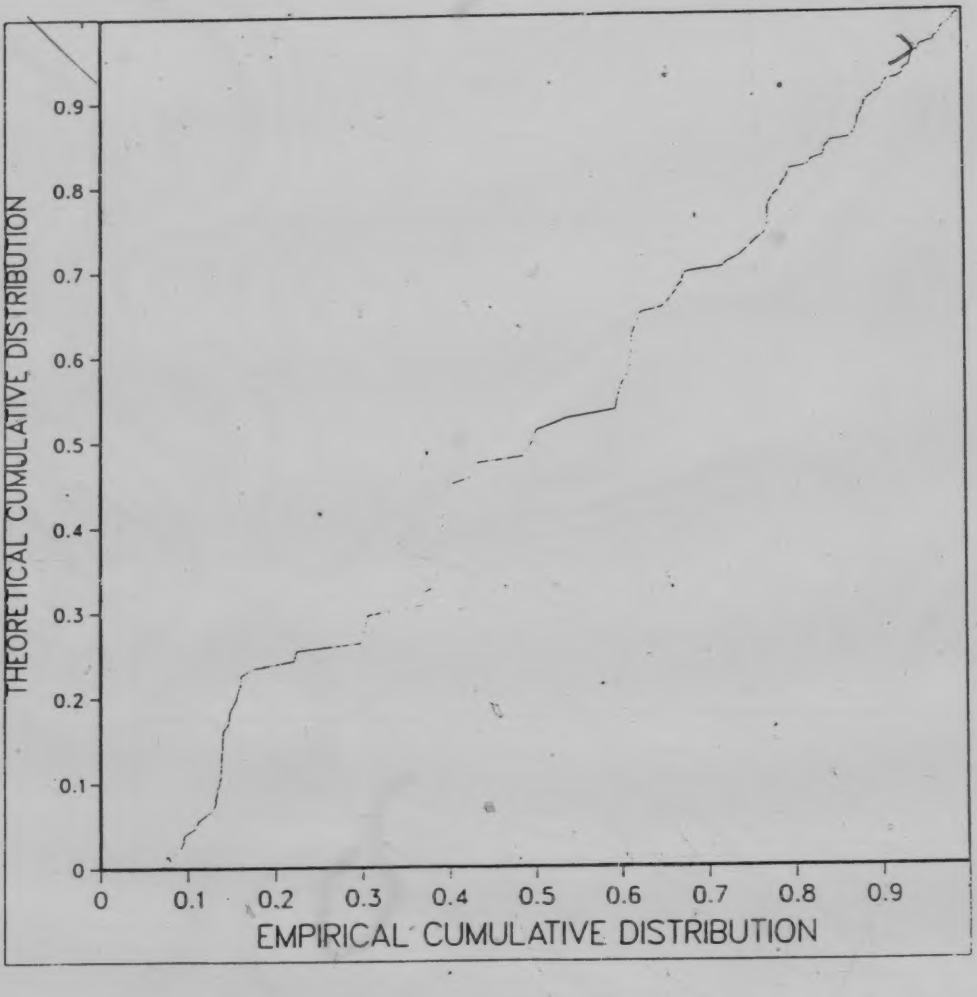

Figure 6.7 ANOVA F with arcsine transformation testing for column effects with no effects present: $p = 0.25$, 10 observations per cell, 2×2 table (case 1.4).

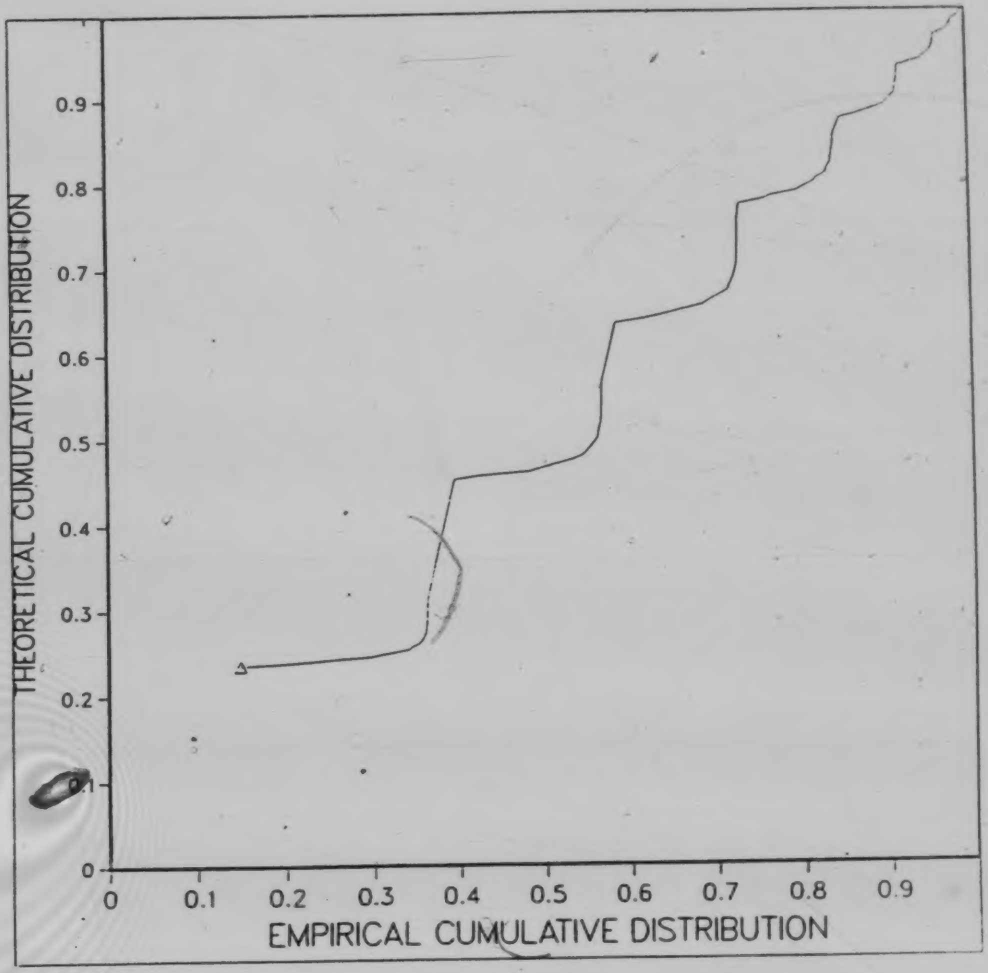

Figure 6.8 ANOVA F testing for column effects with no effects present: $p = 0.5$, 10 observations per cell, 2×2 table (case 1.6).

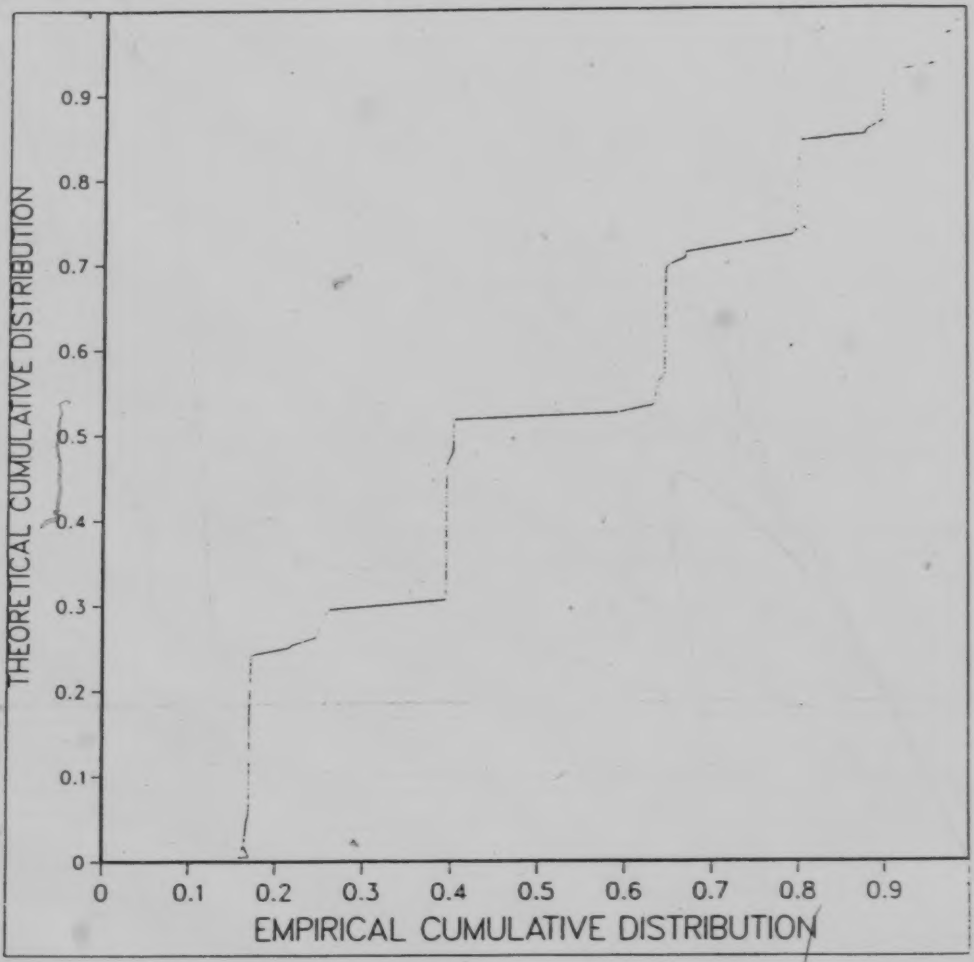

Figure 6.9 ANOVA F testing with arcsine transformation for column effects with no effects present: $p = 0.5$, 10 observations per cell, 2×2 table (case 1.6).

TYPE I ERROR OF ANOVA AND LL AT NOMINAL 5 PERCENT SIGNIFICANCE LEVEL WITH ROW EFFECTS PRESENT

			column			interaction	
n (cell)	**case**	**C_{σ}^{2}**	**ANOVA**	**G^2 forward**	**G^2 backward**	**ANOVA**	**G^2‡**
10	**3.5**	**0.0**	0.0399	0.0015	0.0835	0.0407	0.0900
	3.2	**0.4**	0.0499	0.0117	0.0633	0.0472	0.0669
	3.3	**0.6**	0.0534	0.0540	0.0708	0.0518	0.0425
	3.1	**0.0**	0.0465	0.0384	0.0606	0.0466	0.0734
	3.4	**0.2**	0.0409	0.0854	0.0864	0.0383	0.0840
10†	**3.5**	**0.0**	0.0387	0.0014	0.0144	0.0389	0.0110
	3.2	**0.4**	0.0495	0.0093	0.0311	0.0445	0.0197
	3.3	**0.6**	0.0343	0.0165	0.0234	0.0325	0.0082
	3.1	**0.0**	0.0457	0.0356	0.0415	0.0453	0.0468
	3.4	**0.2**	0.0378	0.0119	0.0127	0.0356	0.0099
40	**3.5**	**0.0**	0.0542	0.0015	0.0574	0.0492	0.0704
	3.2	**0.4**	0.0476	0.0132	0.0504	0.0464	0.0579
	3.3	**0.6**	0.0515	0.0450	0.0553	0.0506	0.0738
	3.1	**0.0**	0.0480	0.0434	0.0503	0.0505	0.0533
	3.4	**0.2**	0.0502	0.0522	0.0530	0.0520	0.0777
40†	**3.5**	**0.0**	0.0538	0.0015	0.0416	0.0505	0.0439
	3.2	**0.4**	0.0481	0.0120	0.0427	0.0451	0.0431
	3.3	**0.6**	0.0548	0.0390	0.0459	0.0530	0.0343
	3.1	**0.0**	0.0461	0.0434	0.0455	0.0485	0.0493
	3.4	**0.2**	0.0535	0.0389	0.0393	0.0551	0.0458

† Results after data transformations: arcsine for ANOVA, addition 0.5 to each cell for LL (loglinear analysis).

‡ Overall goodness of fit statistic for log-linear model that includes all but the highest order term (model of independence).

Table 6.4 Empirical Type I error rates for two dimensional models at nominal 5 percent significance level for null hypotheses of no column and interaction effects while row effects are present

Tables 6.4 6.5 and 6.6 give results for cases where the probabilities of success are not the same for all cells of each configuration. In these cases, row and/or column effects are present by design. The results in the aforementioned tables are Type I error rates based on nominal 5% critical values.

The cases studied here allow for an examination of Type I error when heterogeneity of variance is present. The level of heterogeneity is expressed via C_{σ}^{2}, the coefficient of variation of the cell population variances [RoK77], [KeF71]. Of the cases studied, C_{σ}^{2} ranges from 0.0 to 0.6. The results in [RoK77] indicate that the F-test is robust when C_{σ}^{2} is less than 0.8 and the number of observations per cell is 10 or more. Thus, the maximum value of C_{σ}^{2} in this study, 0.6, is not an extreme. On the other hand, Rogan's results were based on sampling from normal distributions, not binomial distributions. Also, since the variance is a function of the mean in binomial sampling, $\sigma^2 = pqn$, instances where extreme heterogeneity of variance is present are found only where there are large, obvious differences in the cell means (relative to $p = 0.5$). Such situations require that some of the cell probabilities be close to 0 or 1, that is, be at the extreme of the probability scale. Hence, where there is extreme heterogeneity of variance there is also with it the problems caused by the extreme probabilities of the binomial distribution.

EMPIRICAL TYPE I ERROR FOR ANOVA AND LL AT NOMINAL 5 PERCENT LEVEL OF SIGNIFICANCE

n (cell)	case	C_{σ}^{2}	ANOVA	ANOVA†	G^2	G^2†
10	5.4	0.4	0.0553	0.0497	0.0684	0.0203
	5.2	0.4	0.0499	0.0462	0.0742	0.0263
	5.1	0.1	0.0442	0.0430	0.0683	0.0405
	5.3	0.25	0.0396	0.0358	0.0804	0.0096
40	5.4	0.4	0.0488	0.0466	0.0606	0.0439
	5.2	0.4	0.0484	0.0899	0.2197	0.1546
	5.1	0.1	0.0532	0.0499	0.0581	0.0517
	5.3	0.25	0.0507	0.0572	0.0836	0.0452

† Results after data transformations: arcsine for ANOVA, addition 0.5 to each cell for LL (loglinear analysis).

Table 6.5 Empirical Type I error rates (nominal 5 percent level) for true null hypothesis of no interaction effects (additive probability model) for two dimensional cases with both row and column effects present

For cases 4.1, 4.2, and 4.3 entries represent Type I error. Entries represent power for cases 6.1 and 6.2. The results in Tables 6.4, 6.5 and 6.6 show that the arcsine transformation does not consistently improve the empirical to nominal agreement for Type I error. When $n = 10$ the transformation leads more often than not to larger discrepancies from nominal levels than does use of the raw data. When $n = 40$ the discrepancies are reduced for both types of data, neither being discernibly superior to the other.

EMPIRICAL TYPE I ERROR RATES AT NOMINAL 5 PERCENT LEVEL

			tests for column effects			tests for layer effects		
n (cell)	case	C^2_σ	ANOVA	G^2 forward	G^2 backward	ANOVA	G^2 forward	G^2 backward
10	4.1	0.0	0.0530	0.0520	0.0546	0.0547	0.0539	0.0555
	4.2	0.4	0.0484	0.0142	0.0557	0.0464	0.0118	0.0536
	4.3	0.0	0.0490	0.0016	0.0526	0.0516	0.0017	0.0560
10†	4.1	0.0	0.0503	0.0328	0.0459	0.0508	0.0344	0.0467
	4.2	0.4	0.0528	0.0102	0.0321	0.0542	0.0088	0.0302
	4.3	0.0	0.0452	0.0004	0.0160	0.0465	0.0001	0.0187
40	4.1	0.0	0.0457	0.0422	0.0467	0.0472	0.0450	0.0483
	4.2	0.4	0.0490	0.0124	0.0549	0.0459	0.0131	0.0567
	4.3	0.0	0.0495	0.0014	0.0505	0.0522	0.0010	0.0540
40†	4.1	0.0	0.0446	0.0422	0.0431	0.0464	0.0450	0.0458
	4.2	0.4	0.0491	0.0115	0.0476	0.0553	0.0120	0.0482
	4.3	0.0	0.0509	0.0014	0.0406	0.0527	0.0010	0.0416

† Results after data transformations: arcsine for ANOVA, addition 0.5 to each cell for LL (loglinear analysis).

Table 6.6 Empirical Type I error rates for three dimensional cases where a main effect is present.

When both row and column effects are present (Table 6.5) the F-test for interaction effects (none by design in probability model) exhibits peculiar behavior for $n = 40$ when the arcsine transformation is used. In two cases (5.2, 5.3) the empirical level with the transformation is noticeably larger than the nominal level and the level found using the raw scores. These two cases are ones where log-odds interaction effects are present. The levels suggest that non-additivity in the arcsine scale falls between additivity in the raw proportions and non-additivity in the log-odds.

Assessing the effect of heterogeneity of variances and the effectiveness, if any, of the arcsine transformation in improving the F-test is not a straightforward process in this study. The difficulty stems from the fact that larger heterogeneity of variances is not independent of the existence of small minimum cell probabilities in tables. Previous studies [RoK77] [Nor52] have shown that the effect of variance heterogeneity is to inflate the Type I error levels above the nominal levels given normal sampling. This study has shown that when cell probabilities are at the extremes or middle of the 0 to 1 scale, Type I error rates are significantly smaller than the nominal rates when no heterogeneity of variances exists. When heterogeneity of variances is introduced, there might then be a counterbalancing effect in cases where probabilities are extreme and variance heterogeneity is large. That is, small probabilities cause reduced Type I error levels while variance heterogeneity increases Type I error levels. Hence, it seems plausible that given both there may be a cancellation in the discrepancies from the nominal level. The results in Tables 6.4, 6.5, and 6.6 do not show such a counterbalancing; instead, there is an increase in irregular results, especially for the ANOVA using the arcsine transformation. A clear pattern like the one before is not seen with these particular results. This is perhaps evidence that probability size and variance heterogeneity are interacting, but a clear characterization of the interaction is difficult due to the irregular results.

For the case with the largest degree of heterogeneity (case 3.3, Table 6.4) the Type I error is inflated for the raw data for both $n = 10$ and $n = 40$. When the arcsine transformation is applied, the level drops significantly well below nominal for $n = 10$ and rises significantly above nominal for $n = 40$. Case 3.3, Table 6.4, has the probability structure (.05, .05, .25, .25). After application of the arcsine transformation, the underlying cell probabilities appear to have the same effect upon the Type I error levels as they did in the study where no variance heterogeneity was present. For example, for case 3.3 with $n = 10$, the realized level of 0.0343 in the hypothesis of no column effects is −7.14 standard deviations from the nominal 5% level using transformed scores. Looking back to Table 6.3 reveals that for $p = 0.05$ and $p = 0.25$ the corresponding realized levels are −18.1 and 4.0 standard deviations from the nominal 5% level, respectively. The average is −7.05 which is close to that found in case 3.3. Case 3.5 and case 4.3 are also directly comparable to the results in Table 6.3, and the same sort of relationship holds.

When heterogeneity of variances exists and when samples are as small as 10 observations per cell use of the arcsine transformation results in a much more conservative test when using the nominal tables. However, when observations are 40 per cell the transformation gives results almost the same as does an analysis without the transformation (close to nominal), though in several instances it was observed that it significantly increased Type I error.

The results support Lunney's conclusion [Lun70] that the standard fixed-effects ANOVA is adequate when the coefficient of variation is less than 0.8. Furthermore, when sample sizes are as small as 10 observations per cell the standard ANOVA leads to results closer to nominal than does ANOVA using the arcsine transformation. Therefore, when the heterogeneity of variances is not extreme, use of the arcsine transformation is not advised. Cases where extreme heterogeneity of variance is present were not considered in this study, but the behavior of the transformation where cell probabilities are close to 0 or 1 suggest that it is ill-suited unless a very conservative test is desired.

6.3.2. Power

Results for power are found in Tables 6.7, 6.8, 6.10, 6.11, 6.12, 6.13. When $C^2_{\sigma^2}$ is 0.2 or less the arcsine transformation generally causes a slight decrease in power. When greater heterogeneity of variances is present the transformation significantly increases power, but not consistently. The biggest increase in power found in the results occurs when all main effects are present and $C^2_{\sigma^2}$ is 0.4 (cases 5.4 and 5.2 in Tables 6.11 and 6.12). However, the change in power is erratic depending upon sample size and probability structure of the tables. For instance, case 5.4 in Table 6.11 shows the arcsine transformation creates an increase in power of 0.019 at $n = 10$, yet a decrease in power of 0.006 when $n = 40$. Then in case 3.4 of Table 6.9 the transformation leads to a decrease in power of 0.005 at $n = 10$ and an increase of 0.002 at $n = 40$. In general the increases in power are greater than the decreases when heterogeneity of variances is present. In the absence of heterogeneity the transformation always either decreases power or causes no change in power. When the sample size is larger, both transformed and raw data lead to very similar results. This agrees with the results for Type I error.

Table 6.7 presents a comparison of power in the manner recommended in [Mar87]. The column headed with **diff** contains the differences in rejection rates (transformed results minus raw score results). The column headed by **SE** contains the estimated standard deviation of the difference which was calculated using the expression found in [Mar87]. The next column gives for each case the proportion of 10,000 samples where both methods reject the null hypothesis. The final column contains the rejection correlation coefficient for the two calculations of F. This is the correlation between the two arrays formed by scoring 1 if the test rejects and 0 otherwise.

POWER OF ANOVA F AND ANOVA F WITH ARCSINE TRANSFORMATION AT NOMINAL 5 PERCENT LEVEL

n (cell)	case	C_σ^2	ANOVA-T	ANOVA	diff	SE	% both reject	corr
10	3.5	0.0	1.0000	0.9864	0.0136*	0.00116	0.9864	undf
	3.2	0.4	0.9882	0.9894	–0.0012*	0.00037	0.9881	0.94
	3.3	0.6	0.4539	0.4530	0.0009	0.00151	0.4421	0.95
	3.1	0.0	0.2143	0.2202	–0.0059*	0.00093	0.2129	0.97
	3.4	0.2	0.0527	0.0554	–0.0027*	0.00066	0.0519	0.96
40	3.5	0.0	1.0000	1.0000	0.0000	0.00000	1.0000	undf
	3.2	0.4	1.0000	1.0000	0.0000	0.00000	1.0000	undf
	3.3	0.6	0.9645	0.9642	–0.0003	0.00066	0.9622	0.94
	3.1	0.0	0.7165	0.7166	–0.0001	0.00010	0.7165	0.99
	3.4	0.2	0.1380	0.1367	0.0013	0.00147	0.1265	0.91

Entries of **undf** in the corr column represent cases where the correlation coefficient is undefined due to division by zero.
* The difference of the two powers is significant at 5% level.

Table 6.7 Power comparison at nominal 0.05 significance level between ANOVA F for raw scores and ANOVA F after arcsine transformation of scores.

The results show that for $n = 10$, 4 of the 5 cases have a difference that is greater than 2 standard deviations (significant at the 5% level). When $n = 40$ none of the differences are significant. The proportion where both reject is mostly a function of effect size. Looking at this in conjunction with the rejection correlation shows that both methods behave very similarly. Even when both reject infrequently they do so in a highly correlated manner; they are rejecting for the same samples most of the time. Case 3.3 is interesting since it is the case with the largest C_σ^2 that was examined in this study, 0.6. The transformation makes little difference here. This is not unexpected given the results in [Lun70] where it was shown that ANOVA using raw scores is robust when $C_\sigma^2 \leq 0.8$ (when sampling is from normal populations). Based on the power results of this study, the arcsine transformation does not appear to consistently improve the performance of

ANOVA. In some cases it brings noticeable improvements, while in others it causes a reduction in power. This erratic behavior is likely attributable to the effects of small expected counts on the transformation.

			EMPIRICAL POWER AT NOMINAL 5 PERCENT LEVEL					
			tests for column effects			**tests for layer effects**		
n (cell)	**case**	C^2_σ	**ANOVA**	G^2 **forward**	G^2 **backward**	**ANOVA**	G^2 **forward**	G^2 **backward**
10	6.1	0.2	0.4682	0.4559	0.4764	0.4699	0.4582	0.4777
	6.2	0.3	0.9719	0.9692	0.9743	0.5184	0.4494	0.5389
10†	6.1	0.2	0.4680	0.3695	0.4486	0.4660	0.3666	0.4517
	6.2	0.3	0.9728	0.9490	0.9687	0.5359	0.3559	0.4763
40	6.1	0.2	0.9660	0.9588	0.9671	0.9648	0.9559	0.9657
	6.2	0.3	1.0000	1.0000	1.0000	0.9846	0.9702	0.9854
40†	6.1	0.2	0.9652	0.9588	0.9639	0.9631	0.9559	0.9625
	6.2	0.3	1.0000	1.0000	1.0000	0.9878	0.9702	0.9839

† Results after data transformations: arcsine for ANOVA, addition 0.5 to each cell for LL (loglinear analysis).

Table 6.8 Power of ANOVA F and Conditional G^2 for three dimensional cases where all main effects are present.

6.3.3. Further Discussion

The author was not able to locate published Monte Carlo results comparing ANOVA for binomial data with and without the arcsine transformation. Since most thorough textbooks on the subject of experimental design and ANOVA make the recommendation that the arcsine transformation be applied when it is known the data is drawn from binomial populations, it is of some concern that the results of this study suggest that such recommendations are misguided. Therefore, much attention has been directed in an effort to verify and explain the results relating to the arcsine transformation for ANOVA even though it is not the central topic of this thesis.

The increases and decreases in the arcsine transformation results compared to the respective raw score results seem to make a little more sense. The results in a study by Mosteller and Youtz [MoY61] indicate that the variance of the transformed scores is at a relative maximum around $p = 0.2$ for $n = 10$. This means that it is at its greatest positive departure from the asymptotic variance, which is used as the denominator for the F-ratio.[11] As a result, the F-ratio that was calculated in the simulation is slightly greater than what would be expected if the true variance of the transformed population had been used in place of the asymptotic variance estimate.

When cell sample size is large, the asymptotic variance of the transformed scores is approximately the same for all values of p, the probability. Accordingly, an F-ratio which is larger than what would be theoretically expected leads to a higher than nominal Type I error rate. The results of Table 6.3 are consistent with this line of reasoning. For $n = 10$, the results using the arcsine transformation show a higher than nominal rate for probabilities in the vicinity of 0.2, a local maximum for σ_θ^2, the actual variance. For $p \leq 0.1$ the Type I error rate is lower than nominal (true for raw score ANOVA, too) as σ_θ^2 is smaller than the asymptotic variance, decreasing rapidly as p approaches zero. For $p > 0.2$, σ_θ^2 decreases to a local minimum at $p = 0.5$ which is about equivalent to the minimum at $p = 0.1$. The results show that the Type I error rates are about the same for $p = 0.1$ and $p = 0.5$ when the arcsine transformation is employed. Both are conservative but the trends of closeness of agreement to nominal levels differ as α goes from 0.10 to 0.01.

Results for $n = 40$ reveal the same correspondence to the Mosteller-Youtz results. Their results are for $n = 50$ but they are still quite related to the trends exhibited in Table 6.3. The local maximum of σ_θ^2 is at $p = 0.04$. Case 1.1 ($p = 0.05$) of Table 6.3 when $n = 40$ is the only case where Type I error rates clearly exceed the nominal level. The remaining cases all have σ_θ^2 close to, but slightly less than, the asymptotic variance of the transformed scores, with the closest match occurring at $p = 0.1$ which has all empirical rates being within one standard deviation of the nominal rates.

---

[11] Both Box [BHH78] and Mosteller and Tukey [MoT68] comment that use of the sample variance is an alternative that might be preferable since it involves use of more information from the sample.

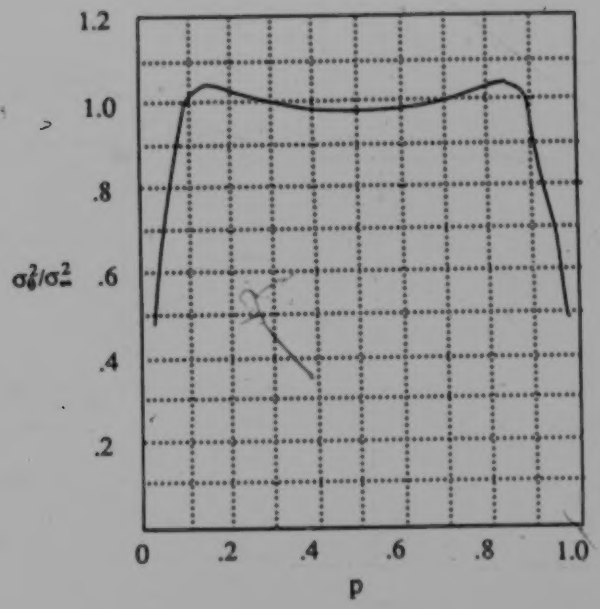

Figure 6.10 Graph of ratio of actual to asymptotic variance, $\sigma_\phi^2/\sigma_\infty^2$, versus binomial probability, p, for $n = 10$ (adapted from Mosteller and Youtz)

The above observations explain the empirical results reasonably well with respect to the Mosteller-Youtz study when viewing the arcsine transformation results exclusively. However, the trends noted for the arcsine transformation results with $n = 10$ are strongly parallel to the results that are arrived at without use of the transformation. The trend is that the discrepancies of the ANOVA F from nominal Type I error rates when plotted against the probabilities, p, follow the same "rabbit ears" shaped curve as that given by Mosteller and Youtz, the plot of the ratio of actual to asymptotic variance versus probability (Figure 6.10 and Figure 6.11). This is the case for both 2 and 3 dimensional configurations examined in this study.

The reason for this has not been clearly established in this study (nor addressed in detail). However, the graphical analysis at $n = 10$ reveals a similarity in trend in that when $p = 0.25$ the cumulative distribution of F is much more in agreement with the nominal distribution than are the cumulatives when $p = 0.1$ and $p = 0.5$.

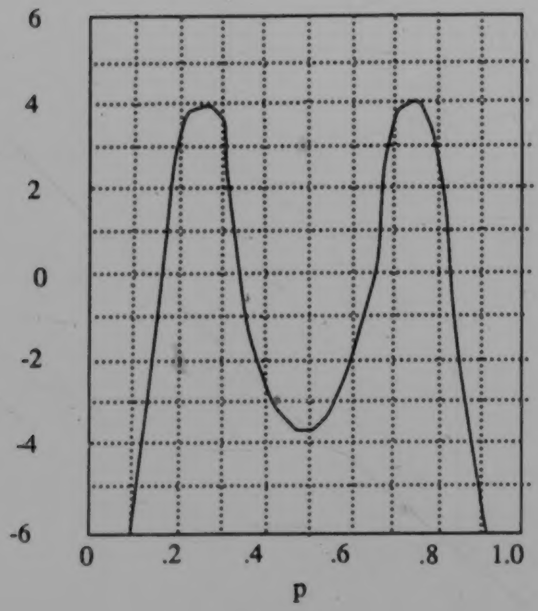

Figure 6.11 Graph of number of standard deviations from nominal 5% level versus binomial probability, *p*, for $n = 10$

Mosteller and Youtz offer no explanation in their paper for the behavior of σ_θ^2. It has not been established in this thesis what causes the behavior of the Type I error discrepancies. What has been shown is that there is a striking similarity in the behavior of Type I error and σ_θ^2 with respect to binomial probabilities when n is small. Whether or not the cause is the same for both (*i.e.*, an underlying property of the binomial distribution), has not been identified in this thesis; however, examining such an assumption would seem to be a reasonable approach to take should one decide to carry out a deeper investigation into the matter.

6.4. Addition of 0.5 to Cell Totals

The small sample performance of the X^2 and G^2 statistics has been thoroughly examined over the last 20 years. It has been clearly established that the performance of each breaks down severely enough to make use of the nominal tables unreliable when there are many small or zero counts. (See Chapter 4 Section 4.1 for more details). The problem has been approached in several ways such as developing new statistics and adjusting the approximating distributions of the usual statistics so that the agreement is better.

Goodman is often credited with making the recommendation of adding 0.5 to each observed cell total in an early paper on loglinear analysis [Goo70]. However, he made the suggestion in relation to calculation of effect parameters in the saturated model with the aim of reducing both the asymptotic bias and standard errors. He did not comment in [Goo70] about the general use of the adjustment, yet it appears that the inclusion of such an option in some popular computer programs for loglinear analysis (*e.g.*, ECTA and BMDP/4) has led many to use the adjustment in the general context of model fitting. The adjustment does have the advantage of eliminating the problem of zeros in the marginal totals (problem for maximum likelihood estimation) and the problem of zeros in elementary cells (problem in definition of G^2). Subsequent to Goodman's paper, several authors have either repeated or cited the recommendation in introductory papers and texts on loglinear analysis [Rey77], [KnB80], [Ken83] [Fox84]. One should note that this adjustment is not the same as the common correction for continuity where 0.5 is subtracted from the difference of expected and observed counts before squaring in the Pearson X^2 statistic.

Intuitively, a uniform increase in each cell total when expected counts are small will result in smaller values of X^2 and G^2 since the effects will be smoothed over. Thus, one should expect a decrease in the number of test statistics that are significant based on critical values of the approximating χ^2 distribution. Type I error should decrease as should power. The point of this part of the study was to determine whether the bias introduced by the adjustment is innocuous or significant. The author of this thesis did not locate any Monte Carlo results studying the effect of the adjustment.[12]

---

[12] There are two papers by Hosmane [Hos86], [Hos87] that investigate a variety of adjustments where a constant is added to cell totals; however, the statistics are also adjusted for asymptotic biases caused by the additions.

STANDARD DEVIATIONS OF X^2 AND G^2 FROM NOMINAL SIGNIFICANCE LEVELS

n (cell)	stat	case	prob	nominal level (α) 0.10	0.05	0.025	0.01
10	X^2	1.2	.1	–27.5	–20.6	–15.3	–9.8
		1.4	.25	–16.5	–14.5	–9.7	–7.3
		1.6	.5	–9.8	–9.1	–6.3	–6.7
		2.1	.1	–28.0	–20.9	–15.0	–9.5
		2.2	.25	–19.6	–15.6	–11.9	–8.7
		2.3	.5	–10.8	–9.9	–7.8	–6.6
	G^2	1.2	.1	–13.1	–10.4	–8.6	–6.2
		1.4	.25	–3.4	–2.3	–2.8	–3.3
		1.6	.5	–1.9	–0.9	–0.6	–0.7
		2.1	.1	–13.5	–10.9	–8.0	–5.0
		2.2	.25	–4.6	–3.4	–3.6	–2.8
		2.3	.5	–1.6	–2.3	–1.9	–1.7
40	X^2	1.2	.1	–27.4	–21.2	–15.1	–9.8
		1.4	.25	–10.9	–12.3	–8.2	–6.7
		1.6	.5	–3.5	–2.6	–2.2	–3.2
		2.1	.1	–30.8	–21.9	–15.8	–10.1
		2.2	.25	–15.0	–13.4	–11.1	–7.8
		2.3	.5	–3.6	–4.0	–2.7	–1.8
	G^2	1.2	.1	–9.1	–7.0	–6.9	–5.5
		1.4	.25	–1.8	0.0	–1.3	–2.3
		1.6	.5	–1.4	–0.5	0.6	0.9
		2.1	.1	–8.6	–7.9	–6.3	–5.4
		2.2	.25	–2.8	–0.4	–0.6	–2.6
		2.3	.5	–0.1	–0.2	–0.1	–0.4

Table 6.9 Standard deviations of G^2 and X^2 from nominal Type I error rates after cell counts have been adjusted by addition of 0.5 to each cell (results without adjustment are in Table 6.14).

The results of this study demonstrate clearly that the adjustment is more harmful than it is useful if one plans to use the nominal χ^2 tables. Table 6.9 gives results for the Type I error study. Tables 6.8, 6.10, 6.11, 6.12, 6.13 display results from the power study. The addition of 0.5 to each cell leads to actual rejection rates that are much lower than the rates for both the X^2 and G^2 statistics without any adjustment. Also, in most cases the adjustment leads to actual levels that are considerably less than the nominal levels.

EMPIRICAL POWER OF ANOVA AND LL AT
NOMINAL 5 PERCENT LEVEL

n (cell)	case	C^2_σ	ANOVA	G^2 forward	G^2 backward	G^2 no effects	X^2 no effects
	3.5	0.0	0.9864	1.0000	1.0000	0.9997	0.9994
	3.2	0.4	0.9894	0.9912	0.9928	0.9658	0.9255
10	3.3	0.6	0.4530	0.4943	0.5064	0.3636	0.0761
	3.1	0.0	0.2202	0.2144	0.2505	0.1818	0.1374
	3.4	0.2	0.0554	0.1107	0.1117	0.0598	0.0061
	3.5	0.0	1.0000	1.0000	1.0000	0.9997	0.9995
	3.2	0.4	0.9882	0.9879	0.9880	0.9487	0.9336
10†	3.3	0.6	0.4539	0.3224	0.3260	0.0997	0.1001
	3.1	0.0	0.2143	0.2073	0.2105	0.1417	0.1066
	3.4	0.2	0.0527	0.0205	0.0212	0.0059	0.0067
	3.5	0.0	1.0000	1.0000	1.0000	1.0000	1.0000
	3.2	0.4	1.0000	1.0000	1.0000	1.0000	1.0000
40	3.3	0.6	0.9645	0.9667	0.9668	0.9070	0.8666
	3.1	0.0	0.7166	0.7164	0.7167	0.5540	0.5491
	3.4	0.2	0.1367	0.1411	0.1417	0.1232	0.0679
	3.5	0.0	1.0000	1.0000	1.0000	1.0000	1.0000
	3.2	0.4	1.0000	1.0000	1.0000	1.0000	1.0000
40†	3.3	0.6	0.9642	0.9608	0.9608	0.8828	0.8715
	3.1	0.0	0.7165	0.7164	0.7164	0.5377	0.5348
	3.4	0.2	0.1380	0.1160	0.1162	0.0738	0.0646

† Results after data transformations: arcsine for ANOVA, addition 0.5 to each cell for LL.

Table 6.10 Power for two dimensional cases with row effects present; null hypothesis of no row effects.

Even when G^2 for raw data has inflated levels (as occurs when expected cell counts are in the interval 1.5–4.0) the adjustment leads to overly conservative results. As expected, the reduced Type I error is matched with a noticeable decrease in power. This decrease in power is found for all cell probabilities and arrangements selected for this study. When the minimum cell expectation (MCE) is small the adjustment leads to especially large Type I discrepancies from nominal performance. This outcome is of note since it occurs in the situation in which one might feel the adjustment would do the most good (*i.e.*, when the MCE is small the occurrance of observed zeros for cell totals is more common).

The poor performance of the adjustment in the presence of small cell expectations is not too surprising. This study as well as others [Odo70], [Lar78], [Rud87] have shown that both G^2 and X^2 do not reject often enough when there are several expected cell counts less than 1.5. When 0.5 is added to each cell there is a further decrease in Type I error.

EMPIRICAL POWER OF ANOVA AND LL AT NOMINAL 5 PERCENT LEVEL

n (cell)	case	C^2_σ	ANOVA row	G^2 forward	G^2 backward	G^2 no effects	X^2 no effects
10	5.4	0.4	0.3660	0.1613	0.3958	0.9854	0.9702
	5.2	0.4	0.3212	0.3100	0.3644	0.5153	0.1831
	5.1	0.1	0.2406	0.2149	0.2761	0.3278	0.2535
	5.3	0.25	0.0556	0.1113	0.1128	0.0662	0.0080
10†	5.4	0.4	0.3854	0.1569	0.2969	0.9767	0.9639
	5.2	0.4	0.3638	0.2718	0.2861	0.2712	0.2304
	5.1	0.1	0.2313	0.2065	0.2251	0.2671	0.2120
	5.3	0.25	0.0532	0.0205	0.0212	0.0084	0.0091
40	5.4	0.4	0.9042	0.7624	0.9066	1.0000	1.0000
	5.2	0.4	0.8714	0.8601	0.8766	0.9822	0.9689
	5.1	0.1	0.7367	0.7248	0.7441	0.8873	0.8854
	5.3	0.25	0.1340	0.1372	0.1392	0.1786	0.0984
40†	5.4	0.4	0.8987	0.7620	0.8976	1.0000	1.0000
	5.2	0.4	0.8970	0.8511	0.8658	0.9784	0.9722
	5.1	0.1	0.7322	0.7247	0.7321	0.8814	0.8796
	5.3	0.25	0.1390	0.1122	0.1127	0.1140	0.0963

† Results after data transformations: arcsine for ANOVA, addition 0.5 to each cell for LL.

Table 6.11 Power for two dimensional cases with row and column effects present; null hypothesis of no effects(LL overall fit), no row effects(ANOVA, conditional G^2).

Thus, even though the adjustment does remedy the problem of encountering zeros for cell totals and zeros in the marginal totals, the bias it introduces increases the bias that is already found when cell totals are small. This discrepancy is large enough that it outweighs the benefits of not having to modify the analysis to deal with zeros. Other techniques exist for handling zeros (*e.g.*, see [Fie77], [Sim85], [LaU84], [Koe86]). These should be considered for use in place of the 0.5 adjustment.

EMPIRICAL POWER OF ANOVA AND LL AT NOMINAL 5 PERCENT LEVEL

n (cell)	case	C_σ^2	ANOVA column	G^2 forward	G^2 backward
10	5.4	0.4	0.9120	0.9874	0.9940
	5.2	0.4	0.3207	0.3098	0.3626
	5.1	0.1	0.2387	0.2126	0.2771
	5.3	0.25	0.0573	0.1094	0.1112
10†	5.4	0.4	0.9901	0.9868	0.9898
	5.2	0.4	0.3617	0.2733	0.2865
	5.1	0.1	0.2287	0.2056	0.2235
	5.3	0.25	0.0540	0.0190	0.0199
40	5.4	0.4	1.0000	1.0000	1.0000
	5.2	0.4	0.8670	0.8564	0.8717
	5.1	0.1	0.7377	0.7217	0.7423
	5.3	0.25	0.1348	0.1377	0.1402
40†	5.4	0.4	1.0000	1.0000	1.0000
	5.2	0.4	0.8915	0.8487	0.8608
	5.1	0.1	0.7322	0.7217	0.7325
	5.3	0.25	0.1423	0.1121	0.1131

† Results after data transformations: arcsine for ANOVA, addition 0.5 to each cell for LL.

Table 6.12 Power for two dimensional cases with row and column effects present; null hypothesis of no column effects.

EMPIRICAL POWER OF ANOVA AND LL AT NOMINAL 5 PERCENT LEVEL

			testing for row effects (present)				
n (cell)	case	C^2_σ	ANOVA	G^2 forward	G^2 backward	G^2 no effects	X^2 no effects
10	4.1	0.0	0.5509	0.4559	0.4563	0.3210	0.2564
	4.2	0.4	1.0000	1.0000	1.0000	0.9990	0.9967
	4.3	0.0	0.9997	1.0000	1.0000	1.0000	1.0000
	6.1	0.2	0.4673	0.4552	0.4764	0.7529	0.6621
	6.2	0.3	0.5247	0.4534	0.5432	0.9868	0.9625
10†	4.1	0.0	0.4186	0.3723	0.3957	0.2329	0.1935
	4.2	0.4	1.0000	1.0000	1.0000	0.9984	0.9977
	4.3	0.0	1.0000	1.0000	1.0000	1.0000	1.0000
	6.1	0.2	0.4638	0.3685	0.4473	0.6624	0.6116
	6.2	0.3	0.5359	0.3627	0.4810	0.9706	0.9625
40	4.1	0.0	0.9515	0.9515	0.9515	0.8195	0.8125
	4.2	0.4	1.0000	1.0000	1.0000	1.0000	1.0000
	4.3	0.0	1.0000	1.0000	1.0000	1.0000	1.0000
	6.1	0.2	0.9661	0.9587	0.9668	0.9999	0.9999
	6.2	0.3	0.9818	0.9682	0.9823	1.0000	1.0000
40†	4.1	0.0	0.9516	0.9515	0.9515	0.8072	0.8002
	4.2	0.4	1.0000	1.0000	1.0000	1.0000	1.0000
	4.3	0.0	1.0000	1.0000	1.0000	1.0000	1.0000
	6.1	0.2	0.9646	0.9587	0.9637	0.9999	0.9999
	6.2	0.3	0.9843	0.9682	0.9807	1.0000	1.0000

† Results after data transformations: arcsine for ANOVA, addition 0.5 to each cell for LL.

Table 6.13 Power for 3 dimensional cases with row,column, layer effects present (cases 6.1,6.2), and cases with only row and column effects (4.1,4.2,4.3): tests for row effects (ANOVA, conditional G^2).

6.5. G^2 Versus X^2

The results in Table 6.14 are consistent with the results obtained in previous studies comparing Type I error of G^2 and X^2 as measures of overall goodness-of-fit. X^2 is consistently closer to the nominal χ^2 distribution than is G^2. In agreement with the results in [Lar78], [Rud87], [GGM70], and [Odo70] this study shows that the likelihood ratio statistic is conservative for small expected counts (counts < 1.5), liberal for moderate expected counts (between 1.5 and 4.0), and reasonably close to nominal (but still not as close as X^2) for expected cell counts greater than 4.0. The boundaries for the intervals just mentioned are those given by Larntz [Lar78]. The granularity of the probabilities chosen for this study were not as fine as those used by Larntz, but interpolation of the results leads to no contradictions. Also, it should be noted that sampling and modelling in this study are based on product multinomial schemes while the other studies examined full multinomial models, though, this should not effect the relative performances of the statistics that are being compared. There is evidence that Type I error increases for G^2 when the only change in the experiment is the dimensionality and, hence, the number of cells. Table 6.14 shows G^2 rejecting more often for the 3 dimensional tables than it does for the two dimensional tables when corresponding cases are compared (correspondence based on cell sample size and cell success probability, *e.g.*, case 1.2 corresponds to case 2.1).

Results comparing the power of G^2 and X^2 can be found in Tables 6.10, 6.11, and 6.13. In this part of the study, the null hypothesis (false) is that no effects are present while the true alternative is that one or more single factor effects (no higher order effects) are present. Thus, depending on the case, the null model is not necessarily a single term (referring to u-terms in the loglinear formulation) away from the true model. The statistics, X^2 and G^2, were calculated for overall goodness-of-fit for the null models and compared to the nominal null distribution to assess power.

Few studies appear to have been carried out that include comparison of power for G^2 and X^2. The one study found in the literature survey that examined power did so via a measure of the strength, "the weighted average of the power, over the parameter space" [CrG82]. The results in this study are consistent with the results of Crook and Good. G^2 is shown to have slightly greater power than X^2 for $\alpha = 0.05$. Regardless of whether the expected counts are small or large, G^2 consistently has higher higher power than X^2. This too is consistent with the findings of Crook and Good. There is some evidence that suggests that the largest differences in power occur when there are expected cell counts less than 1.0. Cases 3.3, and 3.4 of Table 6.10 and cases 5.2, and 5.3 of Table 6.11 include expected cell counts less than 1.0 for $n = 10$. The amounts by which the power of G^2 exceeds the power of X^2 are greatest for these cases. Crook and Good present results that are similar though not so pronounced. They compare G^2 and X^2 with their measure of strength for multinomials with probability vectors of cardinalities 3 and 30 with overall sample sizes of 30 and 10, respectively. In both cases the strength of G^2 exceeds that of X^2. The difference becomes more pronounced in the the latter (small expected counts) of the above two cases examined in their study.

That this discrepancy is largest when expected counts are small is somewhat odd. Since Type I error for G^2 is about the same or less than that of X^2 for small expected counts, one would expect that the power of G^2 would be close to or less than the power of X^2 for the same range of expected counts. The results on power suggest that the opposite is true, that is, G^2 has higher power. (However, since this study was not designed to specifically investigate the matter, a closer examination is warranted before firm conclusions can be made). An explanation of this result follows from the observation in [Lar78] "that very small observed counts increase G^2 and T^2 to a much greater extent than X^2" for moderate expected values (in the range 1.5–4.0). Taking this to mean that differences between small observed counts and expected counts increase G^2 much more than they do X^2, the results for power make sense. In the study of power observed counts of zeros and ones are most likely to occur when expected counts are small. Since the null hypothesis is false there will be differences between observed and expected counts. Therefore, according to the findings of Larntz, G^2 will be increased more than X^2 by those differences. The results on power are consistent with this reasoning.

However, Larntz also showed that observed zeros and ones contribute more to G^2 than to X^2 even when expected frequencies are in the range 1.0–1.5. Hence, to be consistent with the conclusion quoted

STANDARD DEVIATIONS OF X^2 AND G^2 FROM NOMINAL SIGNIFICANCE LEVELS

				nominal level (α)			
n (cell)	stat	case	prob	0.10	0.05	0.025	0.01
10	X^2	1.1	.05	–2.5	–17.9	–9.1	–9.0
		1.2	.1	0.9	–10.4	–1.7	–7.1
		1.3	.175	–6.0	–6.4	–4.0	–3.7
		1.4	.25	–2.4	–4.1	–4.4	–2.2
		1.5	.375	1.8	–0.8	–2.6	–1.5
		1.6	.5	–1.0	–1.5	–2.1	–2.2
	G^2	1.1	.05	–22.9	–16.2	–14.3	–8.9
		1.2	.1	0.1	–0.9	–3.8	–5.7
		1.3	.175	15.8	10.3	8.5	2.3
		1.4	.25	15.7	16.1	11.3	6.5
		1.5	.375	6.5	11.2	9.0	8.6
		1.6	.5	1.2	7.2	5.3	6.5
40	X^2	1.1	.05	–9.2	–5.6	–4.0	–1.3
		1.2	.1	–1.8	–2.7	–3.4	–3.4
		1.3	.175	–1.0	0.2	1.2	–0.5
		1.4	.25	0.2	0.4	–0.3	–1.2
		1.5	.375	–1.9	–1.1	0.1	–0.7
		1.6	.5	0.7	0.7	0.8	0.8
	G^2	1.1	.05	21.1	9.2	0.4	2.5
		1.2	.1	6.4	7.8	5.6	3.7
		1.3	.175	4.4	4.0	4.5	3.6
		1.4	.25	2.4	3.3	2.3	0.9
		1.5	.375	–0.1	0.3	1.8	0.0
		1.6	.5	2.2	1.3	2.1	1.8

Table 6.14 Standard deviations from nominal Type I error rates for X^2 and G^2 as measures of overall goodness-of-fit for null hypothesis (true) of no effects.

above, G^2 should again be more liberal than X^2. However, both the results of Larntz and the results of this study show that G^2 has Type I error about equal to that of X^2 for expected counts in the range 1.0–1.5. Also, results of this study show that G^2 and X^2 have Type I error levels that are lower than nominal and about the same for both when expected counts are smaller than 1.0. Larntz did not offer explanation for this. A possible explanation is that when expected counts (based on marginal totals of the observed counts) are close to zero it is likely that there are several observed zeros in cells since observations must be whole numbers, and these zeros have an effect on G^2 that is different from that which Larntz described for moderate expected cell counts. The estimation procedure will yield positive expected counts for cells with observed zeros if the margins of that cell are not zero. When G^2 is calculated, any cell that has an observed zero contributes zero[13] to the statistic even when the expected count is greater than zero. On the other hand, an observed zero will result in some positive contribution to X^2 in the same situation (see definitions for G^2 and X^2 in Section 3.4 of Chapter 3). In the Monte Carlo simulation, when cell probability parameters are set to be less than 0.1, the occurrence of observed zeros is not rare. Therefore, for cells with observed zeros and nonzero expectations, there will be a contribution to X^2 but not contribution to G^2. Both statistics will be conservative for very small expected cell counts due to zeros in the marginal totals. Furthermore, the typically inflated levels of G^2 (relative to X^2) are reduced to being about the same as those of X^2 due to the increased frequency of observed zeros in cells with nonzero expected counts.

6.6. Comparison of ANOVA and Loglinear Results

A comparison of the results for ANOVA and loglinear analysis is presented in this section. Type I error rates and power comparisons are made for the appropriate statistics. For loglinear modelling the Pearson X^2 statistic and the likelihood ratio statistic G^2 are usually examined to assess the overall fit of the model to the data. Since these are overall goodness-of-fit statistics, they cannot be used to make a direct comparison with the F-tests for individual effects in ANOVA (the null hypotheses are not analogous). On the other hand, the conditional G^2 statistic of loglinear analysis can be used to test hypotheses involving the analogous individual effects parameters of ANOVA models. Therefore, much of the comparison of loglinear and ANOVA modelling is based upon the relative performances of conditional G^2 and F in testing the same hypotheses over the same data. Strictly speaking, the null hypotheses are not identical since the mathematical formulations of the models, and, hence, the parameters, are different. However, analogous hypotheses are the same in that they permit the researcher to identify significant and non-significant contributions of the independent variables (including interactions) to the dependent variable.[14] That is, the hypotheses are the same when they are generalized to be about the presence or lack thereof of "effects" (dependency). For completeness several results for loglinear model overall goodness-of-fit measured by G^2 and X^2 are also included since the statistics are very important to loglinear modelling.

---

[13] This is a different interpretation of an observed zero's contribution to the G^2 statistic than that of Larntz. Larntz interprets the contribution in terms of how much the other observations must be off expectation due to the presence of the zero in a cell.

Overall goodness-of-fit statistics analogous to those of loglinear modelling are not available for regular ANOVA models. However, since ANOVA may be viewed as a special case of multiple regression and correlation analysis (MRC) [Coh82] parallel regression models could be analyzed and the coefficient of determination, R^2, could be used as a goodness-of-fit measure. But, taken alone, R^2 is a descriptive statistic and cannot be used for hypothesis testing in the way X^2, G^2, and F can be. The best analogy to the loglinear hypothesis testing for groups of effects is the F-test for significant increment in R^2 for pairs of multiple regression models [AnB52]. Though this study did not examine multiple regression models, the results of ANOVA hypothesis testing for single effects are identical to those that would be found using MRC [Coh82]. The ANOVA results are adequate for the rudimentary nature of this study, but it should be noted that MRC methods are less restrictive in the assumptions and arguably richer in the results. The advantage of using ANOVA over MRC is the simplicity found in ANOVA for both the calculations and summary results.

6.6.1. Type I Error – No Effects Present

Tables 6.15 through 6.26 give the results for the study of Type I error when no effects are present in 2×2 layouts. The study of 2×2×2 layouts yielded very similar results. The ANOVA F levels are for calculations using the raw data. The headings in the first column of the tables indicate which effects are being tested by the null hypothesis. The last heading, **NULL FIT**, means the statistics measure the overall fit of the null model, *i.e.* ,the model with no effects present. This applies only for G^2 and X^2 of the loglinear modelling since no ANOVA analog was examined in this study.[15]

---

[14] In most cases, the experiments were designed so that the log-odds effects would be present (or absent) to match the presence (or absence) of the additive probability effects in the ANOVA interpretation of the data.

[15] An overall goodness-offit statistic for multiple regression models is described in [Ait74].

STANDARD DEVIATIONS FROM NOMINAL SIGNIFICANCE LEVELS					
		nominal level (α)			
		0.10	0.05	0.025	0.01
ROW	ANOVA F	–14.9	–17.7	–14.5	–9.9
	G^2 forward	29.4	1.4	–8.9	–8.5
	G^2 backward	29.4	1.4	–8.9	–8.5
COLUMN	ANOVA F	–13.2	–16.6	–14.3	–9.9
	G^2 forward	32.2	3.5	–7.5	–8.3
	G^2 backward	32.2	3.5	–7.4	–8.3
INTERACTION	ANOVA F	–15.0	–17.5	–14.3	–9.8
	G^{2*}	8.8	–2.2	–12.2	–8.8
NULL FIT	G^2	–22.9	–16.2	–14.3	–8.9
	X^2	–2.5	–17.9	–9.1	–9.0

* The forward and backward components are the same.

Table 6.15 Type I error results for case 1.1: two factors, n=10 per cell, success probability=0.05 per cell.

STANDARD DEVIATIONS FROM NOMINAL SIGNIFICANCE LEVELS					
		nominal level (α)			
		0.10	0.05	0.025	0.01
ROW	ANOVA F	2.2	–4.7	–7.4	–7.2
	G^2 forward	23.2	16.5	5.2	0.0
	G^2 backward	23.7	16.9	5.4	0.1
COLUMN	ANOVA F	2.1	–4.5	–7.4	–7.6
	G^2 forward	24.1	17.7	6.5	0.9
	G^2 backward	24.3	18.0	6.7	0.9
INTERACTION	ANOVA F	1.5	–4.8	–7.4	–7.4
	G^{2*}	18.8	17.2	–0.3	–1.4
NULL FIT	G^2	0.1	–0.9	–3.8	–5.7
	X^2	1.0	–10.4	–1.7	–7.1

* The forward and backward components are the same.

Table 6.16 Type I error results for case 1.2: two factors, n=10 per cell, success probability=0.1 per cell.

STANDARD DEVIATIONS FROM NOMINAL SIGNIFICANCE LEVELS

		nominal level (α)			
		0.10	**0.05**	**0.025**	**0.01**
ROW	ANOVA F	1.0	0.6	–1.7	–1.4
	G^2 forward	3.0	4.1	6.1	3.6
	G^2 backward	5.0	6.7	6.7	4.4
COLUMN	ANOVA F	1.1	0.4	–2.6	–1.2
	G^2 forward	2.7	3.7	4.7	2.8
	G^2 backward	4.9	6.8	5.3	3.6
INTERACTION	ANOVA F	1.0	–0.7	–2.4	–1.8
	G^{2*}	19.8	20.0	10.6	7.1
NULL FIT	G^2	15.8	10.3	8.5	2.3
	X^2	–6.0	–6.4	–4.0	–3.8

* The forward and backward components are the same.

Table 6.17 Type I error results for case 1.3: two factors, n=10 per cell, success probability=0.175 per cell.

STANDARD DEVIATIONS FROM NOMINAL SIGNIFICANCE LEVELS

		nominal level (α)			
		0.10	**0.05**	**0.025**	**0.01**
ROW	ANOVA F	0.3	0.5	0.9	1.3
	G^2 forward	2.4	1.2	5.6	1.9
	G^2 backward	5.5	5.0	6.0	3.3
COLUMN	ANOVA F	–1.1	–0.5	–1.7	–0.4
	G^2 forward	0.1	–0.1	3.2	0.2
	G^2 backward	3.2	3.5	3.6	2.2
INTERACTION	ANOVA F	–2.2	–0.7	–1.2	0.1
	G^{2*}	13.7	13.0	12.0	10.2
NULL FIT	G^2	15.7	16.1	11.3	6.5
	X^2	–2.4	–4.1	–4.4	–2.2

* The forward and backward components are the same.

Table 6.18 Type I error results for case 1.4: two factors, n=10 per cell, success probability=0.25 per cell.

STANDARD DEVIATIONS FROM NOMINAL SIGNIFICANCE LEVELS		nominal level (α)			
		0.10	0.05	0.025	0.01
ROW	ANOVA F	–3.7	–0.8	1.0	0.4
	G^2 forward	–2.1	–0.1	3.8	2.2
	G^2 backward	6.3	5.2	6.2	3.8
COLUMN	ANOVA F	–3.1	0.2	0.5	0.6
	G^2 forward	–1.3	0.8	4.1	3.4
	G^2 backward	7.1	6.2	6.1	4.6
INTERACTION	ANOVA F	–3.8	–1.1	0.6	–0.5
	G^{2*}	8.5	10.8	10.4	9.4
NULL FIT	G^2	6.5	11.2	9.0	8.6
	X^2	1.8	–0.8	–2.6	–1.5

* The forward and backward components are the same.

Table 6.19 Type I error results for case 1.5: two factors, n=10 per cell, success probability=0.375 per cell.

STANDARD DEVIATIONS FROM NOMINAL SIGNIFICANCE LEVELS		nominal level (α)			
		0.10	0.05	0.025	0.01
ROW	ANOVA F	–6.2	–2.8	–1.5	0.0
	G^2 forward	–6.5	–3.6	1.5	4.3
	G^2 backward	1.4	2.4	8.8	5.7
COLUMN	ANOVA F	–5.3	–2.8	–0.8	0.2
	G^2 forward	–6.5	–3.6	1.5	1.3
	G^2 backward	1.9	2.1	8.1	5.9
INTERACTION	ANOVA F	–4.6	–2.5	–1.9	–0.4
	G^{2*}	5.0	8.9	9.6	6.7
NULL FIT	G^2	1.2	7.2	5.2	6.5
	X^2	–1.0	–1.5	–2.1	–2.2

* The forward and backward components are the same.

Table 6.20 Type I error results for case 1.6: two factors, n=10 per cell, success probability=0.5 per cell.

STANDARD DEVIATIONS FROM NOMINAL SIGNIFICANCE LEVELS		nominal level (α)			
		0.10	**0.05**	**0.025**	**0.01**
ROW	ANOVA F	5.1	–3.3	–3.4	–3.3
	G^2 forward	6.9	8.2	8.6	5.3
	G^2 backward	6.9	8.2	8.6	5.4
COLUMN	ANOVA F	7.6	–1.7	–2.8	–3.0
	G^2 forward	9.3	10.7	9.9	5.0
	G^2 backward	9.3	10.7	10.0	5.1
INTERACTION	ANOVA F	4.2	–4.6	–4.0	–4.4
	G^{2*}	20.9	18.8	8.1	4.0
NULL FIT	G^2	21.1	9.2	0.4	2.5
	X^2	–9.2	–5.6	–4.0	–1.3

* The forward and backward components are the same.

Table 6.21 Type I error results for case 1.1: two factors, n=40 per cell, success probability=0.05 per cell.

STANDARD DEVIATIONS FROM NOMINAL SIGNIFICANCE LEVELS		nominal level (α)			
		0.10	**0.05**	**0.025**	**0.01**
ROW	ANOVA F	–0.5	–0.5	–1.9	–1.4
	G^2 forward	0.6	0.9	1.2	–1.1
	G^2 backward	0.6	1.1	1.6	–0.7
COLUMN	ANOVA F	1.7	–0.2	–1.7	–1.6
	G^2 forward	2.8	1.2	2.4	–1.3
	G^2 backward	2.9	1.4	3.1	–0.9
INTERACTION	ANOVA F	–0.6	–1.6	–1.7	0.1
	G^{2*}	6.2	8.2	6.9	0.3
NULL FIT	G^2	6.4	7.8	5.6	3.7
	X^2	–1.8	–2.7	–3.4	–3.4

* The forward and backward components are the same.

Table 6.22 Type I error results for case 1.2: two factors, n=40 per cell, success probability=0.1 per cell.

STANDARD DEVIATIONS FROM NOMINAL SIGNIFICANCE LEVELS

		nominal level (α)			
		0.10	**0.05**	**0.025**	**0.01**
ROW	ANOVA F	0.2	–0.3	–0.2	0.3
	G^2 forward	1.2	0.9	1.0	0.9
	G^2 backward	1.3	1.1	1.3	1.2
COLUMN	ANOVA F	0.5	–0.1	0.8	–0.5
	G^2 forward	1.7	1.2	1.7	0.5
	G^2 backward	1.7	1.3	2.0	1.2
INTERACTION	ANOVA F	0.4	–0.1	1.3	–0.4
	G^{2*}	3.3	4.6	3.3	2.9
NULL FIT	G^2	4.4	4.0	4.5	3.6
	X^2	–1.9	–1.1	0.1	–0.7

* The forward and backward components are the same.

Table 6.23 Type I error results for case 1.3: two factors, n=40 per cell, success probability=0.175 per cell.

STANDARD DEVIATIONS FROM NOMINAL SIGNIFICANCE LEVELS

		nominal level (α)			
		0.10	**0.05**	**0.025**	**0.01**
ROW	ANOVA F	1.0	0.6	–0.2	–0.1
	G^2 forward	1.3	1.0	1.2	0.7
	G^2 backward	1.7	1.1	1.6	0.8
COLUMN	ANOVA F	–0.4	0.5	0.6	1.4
	G^2 forward	–0.1	1.4	1.8	2.1
	G^2 backward	0.3	1.5	2.0	2.1
INTERACTION	ANOVA F	1.0	1.3	1.7	0.5
	G^{2*}	4.1	2.8	3.3	1.5
NULL FIT	G^2	2.4	3.3	2.3	0.9
	X^2	0.2	0.4	–0.3	–1.2

* The forward and backward components are the same.

Table 6.24 Type I error results for case 1.4: two factors, n=40 per cell, success probability=0.25 per cell.

STANDARD DEVIATIONS FROM NOMINAL SIGNIFICANCE LEVELS					
		nominal level (α)			
		0.10	**0.05**	**0.025**	**0.01**
ROW	ANOVA F	−1.2	−0.3	1.9	0.7
	G^2 forward	0.5	0.6	1.9	0.7
	G^2 backward	0.8	1.0	2.2	0.9
COLUMN	ANOVA F	−2.1	−3.4	−1.5	−0.4
	G^2 forward	−0.6	−2.8	−1.1	−0.5
	G^2 backward	−0.4	−2.2	−0.8	−0.1
INTERACTION	ANOVA F	−2.2	−0.5	1.0	0.0
	G^{2*}	−0.2	1.1	2.2	0.8
NULL FIT	G^2	−0.1	−0.3	1.8	0.7
	X^2	−1.9	−1.1	0.1	−0.7

* The forward and backward components are the same.

Table 6.25 Type I error results for case 1.5: two factors, n=40 per cell, success probability=0.375 per cell.

STANDARD DEVIATIONS FROM NOMINAL SIGNIFICANCE LEVELS					
		nominal level (α)			
		0.10	**0.05**	**0.025**	**0.01**
ROW	ANOVA F	−0.7	−1.2	−1.5	−0.1
	G^2 forward	−0.7	−1.2	−2.0	−0.4
	G^2 backward	−0.7	−1.2	−1.2	0.1
COLUMN	ANOVA F	−1.4	−1.9	−1.0	−0.7
	G^2 forward	−1.4	−2.0	−1.4	−1.3
	G^2 backward	−1.4	−2.0	−0.7	−0.8
INTERACTION	ANOVA F	1.0	1.0	−0.6	0.9
	G^{2*}	1.7	1.4	1.0	2.1
NULL FIT	G^2	2.2	1.3	2.1	1.8
	X^2	0.7	0.7	0.8	0.8

* The forward and backward components are the same.

Table 6.26 Type I error results for case 1.6: two factors, n=40 per cell, success probability=0.5 per cell.

The conditional G^2 statistic used in the tests for individual effects is given for both forward and backward stepwise model selection strategies. Conditional G^2 of the forward procedure is the difference of the overall G^2 for the model with no effects and the overall G^2 for the hierachical model that has highest order term (in logit formulation) being for the effect in question (see Section 3.4 of Chapter 3 for description of conditional G^2). For example, conditional G^2 forward for row effects test is $G^2_{A|B} = G^2_A - G^2_B$ where G^2_A is the likelihood ratio statistic for the model

$$log\,(m_{ijk}) = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} \qquad (model\ A)$$

and G^2_B is for the model

$$log\,(m_{ijk}) = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} \qquad (model\ B)$$

where the subscripts 1, 2, and 3 indicate row, column, and dependent response categories, respectively. The conditional G^2 of the column effects test for the backward procedure is $G^2_{B|D} = G^2_B - G^2_D$ where G^2_D is for the model

$$log\,(m_{ijk}) = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)} \qquad (model\ D)$$

Conditional G^2 for the row effects test using backward selection is $G^2_{C|D} = G^2_C - G^2_D$ where G^2_C is for the model

$$log\,(m_{ijk}) = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{23(jk)} \qquad (model\ C)$$

The issue examined with G^2 forward and G^2 backward is whether or not the use of one or the other makes a difference in the models selected. When there are several terms of the same order there is the issue of deciding which term to test first. This need not be a problem since one could test each term first before deciding which one to include before the next round of selection. In fact, this is what most automated model selection routines do [DiG84]. However, for tables of large dimensionality this becomes a sizable task [EdH87]. Therefore, one specific reason for examining both forward and backward conditional G^2 as described is to see whether there are any outstanding or consistent differences depending upon the choice of starting term when it is one among many of the same order. When the true model is the completely null model, as is the case for the models in this section, the backward elimination strategy may lead to the conditional test of a term using two overspecified models. From a theoretical standpoint conditional G^2 forward should be the statistic compared with the corresponding ANOVA F statistic since it involves the theoretically correct model whereas conditional G^2 backward does not.

The tabled results show that for 40 observations per cell (with cell count expectations ranging from 2 to 20) the ANOVA and loglinear tests for row and column effects are very similar. Overall the ANOVA tests tend to be slightly more conservative than their loglinear analogs. However, neither of the methods gives strong indication of being significantly different from nominal levels at $n = 40$ except when expected cell counts are 2. Overall, the conditional G^2 statistics are slightly liberal while ANOVA F is slightly conservative. For the case with expected counts of 2 ANOVA F is significantly closer to nominal levels than is conditional G^2 forward, yet still not satisfactory. In the case of tests for interaction effects, conditional G^2 forward is in many instances significantly too liberal whereas ANOVA F is close to nominal, though again tending to be slightly conservative. The result for conditional G^2 for interaction effects is not surprising since the statistic was calculated as the difference in G^2's for the model with both row and column effects and the fully saturated model, both models being overspecified. This result is not desirable with respect to stepwise model selection procedures since it means that the interaction term would not be dropped from the model using backward elimination (with forward selection, the model testing concludes before reaching the interaction term in the completely null model cases). On the other hand, the conservative nature of the ANOVA F-test leads to correct rejection of model terms. There is, however, the exception in Table 6.26 where cell count expectation is 20 ($p = 0.5$ and $n = 40$). Here, both conditional G^2 forward and backward for row and column effects are slightly conservative.

As for a distinction between conditional G^2 forward and conditional G^2 backward at $n = 40$ there is little and often no difference. Where differences are present conditional G^2 backward is always more liberal than conditional G^2 forward, as should be expected. However, since these are results where only 2 factors are present the difference is caused by only one extraneous term. In cases where there are multiple factors the difference is more pronounced as is evidenced by comparing conditional G^2 for interaction with conditional G^2 forward for either row or column effects (for example, see Table 6.21).

The study of small sample behavior (n = 10 per cell with expected cell counts of 0.5 to 5.0) gives evidence that ANOVA F approaches the nominal levels more rapidly than does conditional G^2 when going from extremely small to larger expected cell counts. At expected cell counts of 1.75 ANOVA F is acceptable whereas conditional G^2 forward is not. Still, across the range of nominal levels ($\alpha = 0.10, 0.05, 0.025, 0.01$) both statistics exhibit instability in that the direction of deviation from nominal levels changes depending upon the level. This is a result of the discreteness of the actual distributions caused by the small samples. Though not examined in this study, 2-way tables with a greater number of categories per factor, *e.g.*, a 10×7 table, would likely result in a smoothing of the distributions if individual multinomials are fixed in size since the product multinomial sampling would approach the asymptotic model. The same is true for higher dimension as the results for 2×2×2 tables indicate. This, however, is not the case for regular multinomial sampling since as the number of cells increases for a fixed sample size expected cell counts decrease (see [Koe86] for examination of a closely related situation). When expected cell counts are as small as 0.5 neither method gives reliable results. However, there is a difference here with respect to the decisions for model building. The ANOVA F-test is very conservative in the case for all the nominal levels that were investigated while conditional G^2 forward and conditional G^2 backward jump from being liberal to conservative as the level of significance decreases (Table 6.15). The implication is that ANOVA leads to the correct decision whereas loglinear analysis may lead to an erroneous decision depending upon which level of significance is chosen for the test. One should keep in mind, though, that these results are for the case where the underlying model is the completely null model where no effects are present. For cases where effects are present, conditional G^2 forward often exhibits conservative Type I error discrepancies from the nominal 5% level of significance (Tables 6.4, 6.5, 6.6) while conditional G^2 backward is most often liberal. Only the 5% level was examined for these other models, so, unfortunately, a conclusion cannot be made regarding the other levels. One might note, though, that for the completely null model case mentioned above, the conditional G^2 statistics at the 5% nominal level are similarly liberal before jumping to being conservative at the 2.5% level. More extensive results are needed to see if this behavior extends to the other types of null models.

The odd discreteness of ANOVA F at $p = 0.5$ per cell that is described in Section 6.3.1 is paralleled by conditional G^2 forward and conditional G^2 backward. Figures A1.27 and A1.28 give the Type I error plots for both statistics for $n = 10$ and $n = 40$ at $p = 0.5$. Other plots of conditional G^2 show that it behaves very much like ANOVA F. On the other hand, the Pearson X^2 statistic and G^2 for overall goodness-of-fit are somewhat different from ANOVA F and conditional G^2 in that the former two both become more continuous as the cell probabilities approach 0.5 (Figures A1.4, A1.5, A1.6, A1.11, A1.14).

6.6.2. Models With Effects Present – Power and Type I Error

Results for power and Type I error are given in this section for models that include main effects (cases 3.1 through 6.2). The Type I error results are for tests of effects that are not present, either main effects or interaction effects for the examined cases. For the purpose of comparing ANOVA and loglinear analysis, power results are calculated for null hypotheses of single effects being absent when the true alternative is that they are present. Conditional G^2 statistics were studied using both forward and backward selection strategies. However, in this part of the study either conditional G^2 statistic may involve the true model depending upon which effects are present. In the comparisons of power where only one effect is present in the true model both conditional G^2 forward and conditional G^2 backward involve the true model in their calculations. The difference is that conditional G^2 backward is the difference in the true model and a model overspecified by one term whereas conditional G^2 forward is the difference in the true model and the model that is underspecified by the term of interest, the hypothesized null model. Since conditional G^2 forward involves the hypothesized null model in the study of power it most closely resembles the ANOVA F-test in terms of the analogy. Both forward and backward conditional G^2 statistics are compared to ANOVA F in Tables 6.29 and 6.30.

When two effects are present in the true model only conditional G^2 backward involves the true model for tests of main effects, so it most closely resembles the ANOVA F-tests. The conditional G^2 forward statistic will not follow a central χ^2 distribution in this case. For all the cases chosen for this part of the study both forward and backward selection results are tabled alongside the ANOVA F results. One should keep in mind, though, which of the two loglinear analysis results is the more appropriate for comparison with ANOVA F (as discussed above). For several of the cases in the study of power the overall goodness-of-fit statistics G^2 and X^2 were calculated for the completely null model. In the tables where the results for these two statistics are reported they are listed under the heading **no effects**. All ANOVA results mentioned in this section are for ANOVA F without the arcsine transformation and the statistics from loglinear analysis without the addition of 0.5 per cell adjustment.

Below (Tables 6.27 and 6.28) are the logit formulations of the loglinear models examined in this section. The simulation cases along with the conditional G^2 statistics calculated using both forward and backward selection are given. The true models are marked with a [*] in the equations of conditional G^2.

6.6.2.1. Type I Results – Effects Present

Tables 6.4, 6.5, and 6.6 give Type I error results based on nominal 5% significance levels when main effects are present. The Type I error from testing for main effects (Tables 6.4 and 6.6) for ANOVA F generally falls between observed levels of conditional G^2 forward and conditional G^2 backward with $n = 10$ observations per cell. Conditional G^2 backward has levels that are significantly in excess of the nominal rate while conditional G^2 forward is more erratic, at times being excessively liberal or significantly conservative, but always less than conditional G^2 backward. ANOVA F is in most cases closer to the nominal rates than either of the conditional G^2 statistics.

For $n = 40$ observations per cell similar results are found, though they are slightly closer to nominal. The only statistic showing any severe deviations from nominal level for the larger samples is conditional G^2 forward when one effect is present and probabilities are extreme. These are cases 3.2, 3.5, 4.2, 4.3 (Tables 6.4 and 6.6) where conditional G^2 forward is very conservative. Though the discrepancy from the

$[M1]\quad L_{ij} = w^{4}$

$[M2]\quad L_{ij} = w^{4} + w^{14}_{(i)}$

$[M3]\quad L_{ij} = w^{4} + w^{24}_{(j)}$

$[M4]\quad L_{ij} = w^{4} + w^{14}_{(i)} + w^{24}_{(j)}$

$[M5]\quad L_{ij} = w^{4} + w^{14}_{(i)} + w^{24}_{(j)} + w^{124}_{(ij)}$

$[M6]\quad L_{ijk} = w^{5}$

$[M7]\quad L_{ijk} = w^{5} + w^{15}_{(i)}$

$[M8]\quad L_{ijk} = w^{5} + w^{25}_{(j)}$

$[M9]\quad L_{ijk} = w^{5} + w^{35}_{(k)}$

$[M10]\quad L_{ijk} = w^{5} + w^{15}_{(i)} + w^{25}_{(j)}$

$[M11]\quad L_{ijk} = w^{5} + w^{15}_{(i)} + w^{35}_{(k)}$

$[M12]\quad L_{ijk} = w^{5} + w^{25}_{(j)} + w^{35}_{(k)}$

$[M13]\quad L_{ijk} = w^{5} + w^{15}_{(i)} + w^{25}_{(j)} + w^{35}_{(k)}$

Table 6.27 Logit models fitted in Monte Carlo study.

case	effects present	G^2 forward	G^2 backward	effect tested by null hypothesis
1.1-1.6	none	$M1^{*} - M2$	$M2 - M4$	row
		$M1^{*} - M3$	$M3 - M4$	col
		$M4 - M5 = M4$	$M4 - M5 = M4$	int
3.1-3.5	row	$M1 - M2^{*}$	$M4 - M2^{*}$	row
4.1-4.5	row	$M6 - M7^{*}$	$M7^{*} - M10$	row
5.1-5.4	(row,	$M1 - M2$	$M2 - M4^{*}$	row
	col)	$M1 - M3$	$M3 - M4^{*}$	col
		$M4^{*} - M5$	$M4^{*} - M5^{*}$	int
6.1-6.2	(row,	$M6 - M7$	$M12 - M13^{*}$	row
	col,	$M6 - M8$	$M11 - M13^{*}$	col
	layer)	$M6 - M9$	$M10 - M13^{*}$	layer

Table 6.28 Summary of models used in calculations of conditional G^2 forward and conditional G^2 backward.

nominal rate is large, the forward model building process would still lead one to correctly exclude the effect being tested. This is consistent with the other findings on Type I error. Given these Type I error results, use of ANOVA or loglinear forward selection is more reliable than use of loglinear backward elimination. Either of the former two methods leads to the correct model building decisions for these cases.

When the tests are for first order interaction effects the pattern of Type I error rates is fairly regular for the test statistics (Tables 6.4 and 6.5). In this case, the G^2 statistic is the usual overall goodness-of-fit statistic for the model of independence.[16] This statistic is liberal except when cell counts are very small. The ANOVA F-test is closer to nominal and tends to be somewhat conservative for small cell counts. These results are consistent with those in Section 6.3.1 of this chapter (though, ANOVA F is significantly liberal in a few cases). As for model building the two methodologies in general may lead to different decisions. This is a result of the different theory underlying the two modelling techniques. ANOVA models are additive in the probabilities whereas logit models are multiplicative in the odds of the success to failure probabilities. Therefore, log-odds effects may be present in the logit models when additive probability effects are absent in the ANOVA models, and *vice versa*. Chapter 4 Section 4.3 gives examples of this in cases taken from [Mag78]. In this study cases 5.2 and 5.3 in Table 6.5 are instances for where log-odds interaction effects are present while additive probability effects are absent. Accordingly, the Type I error (assuming additive probability effects) is very high for the loglinear G^2 statistic.

When effects are the same (*i.e.*, present or absent) in both model formulations ANOVA F is usually closer to nominal Type I error levels than either of the loglinear overall goodness-of-fit statistics G^2 and X^2. Taken alone overall G^2 gives results that are much less reliable than those of ANOVA F. The discrepancies from the nominal level are quite erratic (Section 6.3.1) depending on size of cell counts. Also, X^2 is more stable than G^2, but it is not as close to nominal as is ANOVA F (Section 6.5). In addition, the conditional G^2 statistics in general tend to be a bit further from nominal than ANOVA F in terms of Type I error when the presence of effects is the same for both additive probabilities and log-odds effects models. Hence, in terms of Type I error the loglinear statistics are not as close to nominal as ANOVA F is.

---

[16] However, in this case it is not measuring the fit of the true model. Instead it is measuring an oversaturated one. Tables 6.27 and 6.28 show that it is the same as

the conditional G^2 testing for the interaction effect since $G^2_{4|5} = G^2_4 - G^2_5$ where $M5$ is the fully saturated model (so $G^2_5 = 0$).

6.6.2.2. Power

Of course, the preceding comments are incomplete without consideration of the relative power of the two methods. Tables 6.29 and 6.30 give results comparing ANOVA F to conditional G^2 forward and G^2 backward, respectively, for the null hypothesis of no row effects when the true model includes row effects. Table 6.31 gives the same results for comparing conditional G^2 forward to conditional G^2 backward. As indicated by the last column in Table 6.29 both ANOVA F and conditional G^2 forward are highly correlated in their rejections of the null hypothesis, even when the effects that are present are very small. However, conditional G^2 forward has significantly higher power than ANOVA F for $n = 10$. At $n = 40$ conditional G^2 forward again has higher or equal power to that of ANOVA F. In almost all cases conditional G^2 backward has higher power than conditional G^2 forward (not surprising as it often has greater Type I error).

Examination of Table 6.31 shows that conditional G^2 forward and G^2 backward are very correlated in their rejections of the null hypothesis. For the small samples, conditional G^2 backward has higher power than conditional G^2 forward, but the two are practically the same when sample size is increased to $n = 40$. Again, this suggests that the backward model selection strategy is not identical to forward selection, at least for this type of model.

Tables 6.6, 6.8, 6.10, 6.11, 6.12, 6.13 give additional results for the study of power. In tests for individual main effects conditional G^2 backward has the highest power. ANOVA F has higher power than conditional G^2 forward only in a few instances (no discernible pattern); otherwise, conditional G^2 forward has greater power. As expected, increased sample size significantly increases the power of each statistic. When all main effects are present the results are similar except that ANOVA F has greater power more often than conditional G^2 forward. This is likely due to the fact that for these cases conditional G^2 forward does not involve the true model in its calculation. (See cases 5.1–6.2 in Table 6.28).

POWER OF ANOVA F AND CONDITIONAL G^2 FORWARD AT NOMINAL 5 PERCENT LEVEL

n (cell)	case	C_σ^2	ANOVA	G^2 forward	diff	SE	% both reject	corr
10	3.5	0.0	0.9864	1.0000	–0.0136*	0.00116	0.9864	undf
	3.2	0.4	0.9894	0.9912	–0.0018*	0.00049	0.9891	0.88
	3.3	0.6	0.4530	0.4943	–0.0413*	0.00219	0.4488	0.90
	3.1	0.0	0.2202	0.2144	–0.0058*	0.00115	0.2107	0.96
	3.4	0.2	0.0554	0.1107	–0.0553*	0.00233	0.0544	0.67
40	3.5	0.0	1.0000	1.0000	0.0000	0.00000	1.0000	undf
	3.2	0.4	1.0000	1.0000	0.0000	0.00000	1.0000	undf
	3.3	0.6	0.9645	0.9667	–0.0023*	0.00048	0.9645	0.97
	3.1	0.0	0.7166	0.7164	–0.0002	0.00014	0.7164	0.99
	3.4	0.2	0.1367	0.1411	–0.0044*	0.00068	0.1366	0.98

Entries of **undf** in the **corr** column represent cases where the correlation coefficient is undefined due to division by zero.

* The difference in the two powers is significant at the 5% level.

Table 6.29 Power comparison at nominal 0.05 significance level between ANOVA F for raw scores and conditional G^2 forward.

POWER OF ANOVA F AND CONDITIONAL G^2 BACKWARD AT NOMINAL 5 PERCENT LEVEL

n (cell)	case	C_σ^2	ANOVA	G^2 backward	diff	SE	% both reject	corr
10	3.5	0.0	0.9864	1.0000	–0.0136*	0.00116	0.9864	undf
	3.2	0.4	0.9894	0.9928	–0.0034*	0.00060	0.9893	0.81
	3.3	0.6	0.4530	0.5064	–0.0534*	0.00230	0.4519	0.89
	3.1	0.0	0.2202	0.2505	–0.0303*	0.00190	0.2168	0.90
	3.4	0.2	0.0554	0.1117	–0.0563*	0.00233	0.0549	0.68
40	3.5	0.0	1.0000	1.0000	0.0000	0.00000	1.0000	undf
	3.2	0.4	1.0000	1.0000	0.0000	0.00000	1.0000	undf
	3.3	0.6	0.9645	0.9668	–0.0022*	0.00047	0.9645	0.97
	3.1	0.0	0.7166	0.7167	–0.0001	0.00017	0.7165	0.99
	3.4	0.2	0.1367	0.1417	–0.0050*	0.00072	0.1366	0.98

Entries of **undf** in the **corr** column represent cases where the correlation coefficient is undefined due to division by zero.

* The difference in the two powers is significant a the 5% level.

Table 6.30 Power: ANOVA F and G^2 backward.

POWER OF CONDITIONAL G^2
AT NOMINAL 5 PERCENT LEVEL

n (cell)	case	C_{σ}^{2}	G^2 forward	G^2 backward	diff	SE	% both reject	corr
	3.5	0.0	1.0000	1.0000	0.0000	0.00000	1.0000	undf
	3.2	0.4	0.9912	0.9928	–0.0016*	0.00040	0.9912	0.90
10	3.3	0.6	0.4943	0.5064	–0.0121*	0.00109	0.4943	0.98
	3.1	0.0	0.2144	0.2505	–0.0361*	0.00187	0.2144	0.90
	3.4	0.2	0.1107	0.1117	–0.0010*	0.00032	0.1107	0.99
	3.5	0.0	1.0000	1.0000	0.0000	0.00000	1.0000	undf
	3.2	0.4	1.0000	1.0000	0.0000	0.00000	1.0000	undf
40	3.3	0.6	0.9667	0.9668	–0.0001	0.00010	0.9667	0.99
	3.1	0.0	0.7164	0.7167	–0.0003	0.00017	0.7164	0.99
	3.4	0.2	0.1411	0.1417	–0.0006*	0.00025	0.1411	0.99

Entries of **undf** in the **corr** column represent cases where the correlation coefficient is undefined due to division by zero.
* The difference is significant at the 5% level.

Table 6.31 Power comparison at nominal 0.05 significance level between conditional G^2 statistics given by the forward and backward stepwise model selection strategies of loglinear analysis.

Tables 6.10, 6.11, and 6.13 include power results for X^2 and G^2 for measuring overall goodness-of-fit for the completely null model (*i.e.*, the model with no effects present). Interestingly, these statistics both have lower power than ANOVA F and the conditional G^2 statistics when only one effect is present in the true model (results in Tables 6.10 and 6.11). When all main effects are present X^2 and G^2 have higher power than the individual effects tests (Tables 6.11 and 6.13). The latter situation is not a surprise since the hypotheses being tested are not the same. X^2 and G^2 are testing a null model which deviates more from the true model. Again, these results indicate that use of G^2or X^2 alone for model selection is not as reliable as using the stepwise procedures or ANOVA.[17]

---

[17] However, one must keep in mind that results involving variable interactions may differ for the two models since ANOVA is additive in probabilities while the logit models are for odds and therefore multiplicative.

6.7. Other Measures

In addition to power and Type I error results the rank correlation coefficient and Kolmogorov-Smirnov statistic were examined. The rank correlation coefficient was calculated for various pairings of the loglinear and ANOVA empirical distributions. The Kolomogorov-Smirnov statistic was used to compare the empirical cumulative to nominal cumulative distributions. Both of these proved to be of little use in identifying trends and patterns in the results. For the most part, the Kolmogorov-Smirnov statistic was almost always significant in the upper tail probabilities. This is most likely a result of the large sample size of 10,000. In a sample this large, the critical value for $\alpha = 0.05$ is 0.0136. This is quite small, and given the discreteness of the distributions being compared, a difference of that size or greater does not seem unlikely. Also, when the theoretical cumulative probability is 0.5 the standard deviation of the empirical value is $0.005 = [\,(0.5)(0.5)/10{,}000\,]^{1/2}$.

Results comparing empirical to nominal distribution using the Kolmogorov-Smirnov test do reveal that the statistic is generally smaller for ANOVA using transformed data than for ANOVA using raw data. Also, the statistic is smaller for 3 dimensional cases than it is for 2 dimensional cases. X^2 generally has the smallest Kolmogorov-Smirnov value out of all the statistics across all cases. However, these comments are based upon using the Kolmogorov-Smirnov value as a descriptive statistic since in most cases it was significant (with the exception of several of the 3 dimensional cases at $n = 40$).

The results for the rank correlation coefficient that is described in Section 5.5 of Chapter 5 are not very interesting. For every comparison of pairs of empirical distributions (with same data set) the correlation is extremely high (above 0.97). This result means that all the statistics that were examined follow very similar rankings of empirical values. This was basically anticipated at the outset of this study since the initial assumption was that both loglinear analysis and ANOVA are viable methods for analyzing the same Bernoulli data.

Chapter 7

Conclusion

The main result of this study is that ANOVA and loglinear analysis (logit) perform quite closely to one another when the presence (or absence) of differrences in log-odds treatment effects and additive probability treatment effects coincides. For applied practitioners this result is encouraging but somewhat limitted since the aforementioned case is not the usual state of data. Clearly, when they do not coincide the methods give different results since the modelling of the data is different. In cases of analogous interpretations (*e.g.*, no differences in effects present or only differing main effects present) the results suggest that ANOVA is the stabler (or more consistent) of the two methods, but only slightly so. ANOVA F is on average closer to nominal levels for Type I error than the conditional likelihood ratio statistics of loglinear analysis (which tend to be liberal). Not surprising, the conditional likelihood ratio tests have higher power than ANOVA F. The ANOVA F test does have the advantage over the likelihood ratio, however, since it does not exhibit the erratic behavior of the likelihood ratio for moderate and small cell counts. Also, in loglinear modelling model selection strategies are necessary. The results examining stepwise forward and stepwise backward selection strategies demonstrate that neither is consistently better than the other. In fact, in some cases one might perform quite poorly while the other does reasonably well (as good as ANOVA F). However, in applied situations the experimenter cannot, in general, know which strategy will be best.

In terms of small sample behavior, both methods are about equal. Though neither test is really reliable in this case, ANOVA F is superior to the conditional G^2 when probabilities are extreme and observed counts are small. This is yet another example of the more general case of the overall goodness-of-fit G^2 breaking down for small samples. The results for the Pearson $\mathbf{X}^2$ show that it is close in performance to ANOVA F, but, direct comparisons are not tenable since they do not test the same (or analogous) hypotheses. As for the effect of the product multinomial sampling scheme upon the relative performances of G^2 and $\mathbf{X}^2$, none was noted as expected. The results agree closely with those of other studies using full multinomial sampling.

The results involving the use of the arcsine transformation for ANOVA are somewhat unexpected. Overall, the use of the transformation proved to have little or no effect. For Type I error it exaggerated the discrepancies of the ANOVA F calculated from the raw data (conservative or liberal depending upon the nominal level chosen). The transformation did bring an increase in power in some cases, but this was erratic. Heterogeneity of variances caused by the Bernoulli-type sampling was apparently not large enough to have a significant influence on the ANOVA tests. Still, the fact that the arcsine transformation had little positive effect and at times appeared to be disruptive is disturbing due to its wide acceptance in past and current texts about ANOVA and experimental design. One might note, however, that the transformation introduces a new scale that is between the additive probability and log-odds scales (evidenced by the cumulative plots [Cox70]), and therefore some results will differ since additivity in one scale does not guarantee additivity in the others (see later part of Section 6.3.1). A correlation in the arcsine transformation results and the relationship of the asymptotic variance of the transformed counts to binomial probabilities was noted. This result does not explain the behavior of the transformation, but it does suggest that it is tied to binomial probabilities and sample size in the same way as the asymptotic variance is.

The addition of 0.5 to all elementary cells of a contingency table was shown to cause very large discrepancies in the test statistics from the nominal levels (greatly reduces Type I error below nominal levels). The addition has particularly strong negative effects on loglinear analysis when expected cell counts are very small and the likelihood of observing many zeros for cell totals is high. On the basis of these findings use of such an adjustment is strongly discouraged.

As a final note, cumulative plots of the empirical to nominal distributions showed that ANOVA F at $p = 0.5$ is clearly discrete, more than for $p = .25$ and $p = .175$. This result is odd since one might assume that discreteness should disappear at $p = .5$ since the binomial more closely resembles the normal distribution at this point. The result may be due to the sampling scheme, but how, if it is at all, has not been established in this thesis.

References

[Ait74] M. A. Aitkin (1974), Simultaneous inference and the choice of variable subsets in multiple regression, *Technometrics* **16**, 221-227.

[AnB52] R. L. Anderson and T. A. Bancroft (1952), *Statistical theory in research*, McGraw-Hill, New York.

[Bar47] M. S. Bartlett (1947), The use of transformations, *Biometrics* **3**, 39-52.

[Bir63] M. W. Birch (1963), Maximum likelihood in three-way contingency tables, *Journal of the Royal Statistical Society, Series B* **25**, 220-233.

[BFH75] Yvonne M. M. Bishop, Stephen E. Fienberg and Paul W. Holland (1975), *Discrete Multivariate Analysis: Theory and Practice*, MIT Press, Cambridge [Written with collaboration of Richard J. Light and Frederick Mosteller].

[BHH78] George E. P. Box, W. Hunter and J. Hunter (1978), *Statistics for experimenters: an introduction to design, data analysis, and model building*, Wiley, New York.

[Coc40] W. G. Cochran (1940), The analysis of variance when experimental errors follow the poisson or binomial laws, *Annals of Mathematical Statistics* **11**, 335-347.

[Coh82] Jocab Cohen (1982), New-look" multiple regression/correlation analysis and the analysis of variance/covariance, *Statistical and methodological issues in psychology and social sciences research*, Gideon Keren (ed.), Lawrence Erlbaum Associates, Hillsdale, New Jersey.

[Cox70] D. R. Cox (1970), *The analysis of binary data*, Methuen, London.

[CrG82] James Flinn Crook and Irving John Good (1982), The powers and strengths of tests for multinomials and contingency tables, *Journal of the American Statistical Association* 77(380), 793-802.

[Cur43] J. H. Curtiss (1943), On transformations used in the analysis of variance, *Annals of Mathematical Statistics* **14**, 107-122.

[DeS40] W. E. Deming and F. F. Stephen (1940), On a least squares adjustment of a sampled frequency table when the expected marginal totals are known, *Annals of Mathematical Statistics* **11**, 427-444.

[DiG84] William R. Dillon and Matthew Goldstein (1984), *Multivariate analysis- methods and applications*, John Wiley, New York.

[Dun75] O. D. Duncan (1975), *Introduction to structural equation models* , Academic Press, New York.

[DuC87] O. J. Dunn and V. A. Clark (1987), *Applied statistics: analysis of variance and regression*, John Wiley & Sons, New York.

[EdH87] D. Edwards and T. Havranek (1987), A fast model selection procedure for large families of models, *Journal of the American Statistical Association* **82**(397), 205-213.

[Fel57] W. Feller (1957), *An introduction to probability theory and its applications, volume I, 2nd ed.*, Wiley, New York.

[Fie77] Stephen E. Fienberg (1977), *The Analysis of Cross-Classified Categorical Data*, MIT Press, Cambirdge.

[Fie79] S. E. Fienberg (1979), the use of chi-squared statistics for categorical data problems, *Journal of the Royal Statistical Society, Series B* **41**, 54-64.

[Fis44] R. A. Fisher (1944), *Statistical methods for research workers, 9th ed.*, Oliver and Boyd, Edinburgh.

[Fox84] John Fox (1984), *Linear statistical models and related methods with applications to social research*, Wiley, New York.

[FrT50] Murray F. Freeman and John W. Tukey (1950), Transformations related to the angular and the square root, *Annals of Mathematical Statistics* **21**, 607-611.

[GGM70] Irving John Good, T. N. Gover and G. J. Mitchell (1970), Exact distributions for chi-squared and for the likelihood-ratio statistic for the equiprobable multinomial distribution, *Journal of the American Statistical Association* **65**(329), 267-283.

[GoK54] Leo A. Goodman and W. H. Kruskal (1954), Measures of association for cross-classifications, *Journal of the American Statistical Association* **49**, 732-764.

[GoK59] Leo A. Goodman and W. H. Kruskal (1959), Measures for association for cross-classifications, II, *Journal of the American Statistical Association* **54**, 123-163.

[GoK63] Leo A. Goodman and W. H. Kruskal (1963), Measures for association for cross-classifications, III, *Journal of the American Statistical Association* **58**, 310-364.

[Goo70] Leo A. Goodman (1970), The multivariate analysis of qualitative data: interactions among multiple classifications, *Journal of the American Statistical Association* **65**(329), 226-256.

[Goo78] Leo A. Goodman (1978), *Analyzing qualitative/categorical data*, Abt Books, Cambridge, Mass..

[GSK69] J. E. Grizzle, C. F. Starmer and G. G. Koch (1969), Analysis of categorical data by linear models, *Biometrics* **25**, 489-504.

[Hab74] S. J. Haberman (1974), *The analysis of frequency data*, University of Chicago Press, Chicago.

[Hab78] S. J. Haberman (1978), *Analysis of Qualitative Data, Volume 1*, vol. 1, Academic Press, New York.

[Hab79] S. J. Haberman (1979), *Analysis of Qualitative Data, Volume* 2, vol. 2, Academic Press, New York.

[Hos86] B. S. Hosmane (1986), Improved likelihood ratio tests and Pearson chi-square test for independence in two dimensional contingency tables, *Communications in Statistics- Theory and Methods* **15**(6), 1875-1888.

[Hos87] B. S. Hosmane (1987), Improved likelihood ratio test for multinomial goodness of fit, *Communications in Statistics- Theory and Methods* **16**, 3185-3198.

[HsF69] Tse-Chi Hsu and Leonard S. Feldt (1969), The effect of limitations on the number of criterion score values on the significance level of the F-test, *American Educational Research Journal* **6**(4), 515-527.

[KeF71] Oscar Kempthorne and L. Folks (1971), *Probability, statistics, and data analysis*, Iowa State University Press, Ames, Iowa.

[Ken83] John J. Kennedy (1983), *Analyzing Qualitative Data*, Praeger Publishers, New York.

[Kir82] Roger E. Kirk (1982), *Experimental design: procedures for the behavioral sciences, 2nd ed.*, Brooks/Cole, Monterey, Calif..

[KnB80] D. Knoke and P. J. Burke (1980), Log-linear models, *Sage University Paper Series On Qualitative Applications In The Social Sciences, series no. 07-020*, Sage Publications, Beverly Hills, Ca..

[KoL80] K. J. Koehler and K. Larntz (1980), An assessment of several asymptotic approximations for sparse multinomials, *Journal of the American Statistical Association* **75**, 336-344.

[Koe86] K. J. Koehler (1986), Goodness-of-fit tests for log-linear models in sparse contingency tables, *Journal of the American Statistical Association* **81**(394), 483-493.

[Lan51] H. O. Lancaster (1951), Complex contingency tables treated by partition of chi-squared , *Journal of the Royal Statistical Society, Series B* **13**, 242-249.

[Lar78] Kinley Larntz (1978), Small-sample comparisons of exact levels for chi-squared, *Journal of the American Statistical Association* **73**(362), 253-263.

[LaU84] H. B. Lawal and G. J. G. Upton (1984), On the use of chi-squared as a test of independence in contingency tables with small cell expectations, *Australian Journal of Statistics* **26**, 75-85.

[LiM71] J. Richard Light and Barry H. Margolin (1971), An analysis of variance for categorical data, *Journal of the American Statistical Association* **66**(335), 534-544.

[Lin53] E. F. Lindquist (1953), *Design and analysis of experiments in psychology and education*, Houghton Mifflin Company, Boston/New York.

[Lun70] Gerald H. Lunney (1970), Using analysis of variance with a dichotomous dependent variable: an empirical study, *Journal of Educational Measurement* **7**(4), 263-269.

[Mag78] Jay Magidson (1978), An illustrative comparison of Goodman's approach to logit analysis with dummy variable regression analysis [In L.A. Goodman, Analyzing Qualitative/Categorical Data. Cambridge, Mass.: Abt Books, 1978.].

[Mar87] E. Markowski (1987), Accurate estimation of standard errors when comparing simulated power functions, *Journal of Statistical Computation and Simulation* **27**, 269-275.

[MoY61] Frederick Mosteller and Cleo Youtz (1961), Tables of the Freeman-Tukey transformations for the binomial and Poisson distributions, *Biometrika* **48**, 433-440.

[MoT68] F. Mosteller and J. Tukey (1968), Data analysis, including statistics, *Handbook of social psychology*, vol. 2, L. Gardner and E. Aronson (eds.), Addison-Wesley, Reading, Mass., 80-203.

[Nor52] D. W. Norton (1952), An empirical investigation of some effects of non-normality and heterogeneity on the F-distribution, Ph.D. dissertation, Department of Education, State University of Iowa.

[Odo70] C. L. Odoroff (1970), A comparison of minimum logit chi-square estimation and maximum likelihood estimation in 2 × 2 × 2 and 3 × 2 × 2 contingency tables: tests for interaction, *Journal of the American Statistical Association* **65**, 1627-1631.

[Pea00] K. Pearson (1900), On a criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can reasonably be supposed to have arisen from random sampling, *Philosophical Magazine* **50**, 157-175.

[Pla74] R. L. Plackett (1974), *The analysis of categorical data*, Griffin, London.

[Rey77] H. T. Reynolds (1977), Analysis of nominal data, *Quantitative applications in the social sciences*, Sage Publications, Beverly Hills, Ca..

[RoK77] Joanne C. Rogan and H. J. Keselman (1977), Is the ANOVA F-test robust to variance heterogeneity when sample sizes are equal?: an investigation via a coefficient of variation, *American Educational Research Journal* **14**(4), 493-498.

[RoK56] S. N. Roy and M. A. Kastenbaum (1956), On the hypothesis of no "interaction" in a multi-way contingency table, *Annals of Mathematical Statistics* **27**, 749-757.

[Rud87] T. Rudas (1987), Testing goodness-of-fit of log-linear models based on small samples: a Monte Carlo study, *Goodness-Of-Fit*, K. Sarkadi, P. Revesz and P. K. Sen (eds.), North-Holland Publishing Company, Amsterdam.

[Sch59] H. Scheffe (1959), *The analysis of variance*, Wiley, New York.

[Sim85] Jeffrey S. Simonoff (1985), An improved goodness-of-fit statistic for sparse multinomials, *Journal of the American Statistical Association* **80**(391), 671-677.

[ThR87] D. R. Thomas and J. N. K. Rao (1987), Small-sample comparisons of level and power for simple goodness-of-fit statistics under cluster sampling, *Journal of the American Statistcal Association* **82**(398), 630-636.

[WCT86] Rand R. Wilcox, Ventura L. Charlin and Karen L. Thompson (1986), New Monte Carlo results on the robustness of the ANOVA F, W, and the F* statistics, *Communications in Statistics-Simulation* **15**(4), 933-943.

[Yul00] G. U. Yule (1900), On the association of attributes in statistics, *Philosophical Transactions of the Royal Society, Series A* **194**, 257-319.

Appendix A1

Graphs

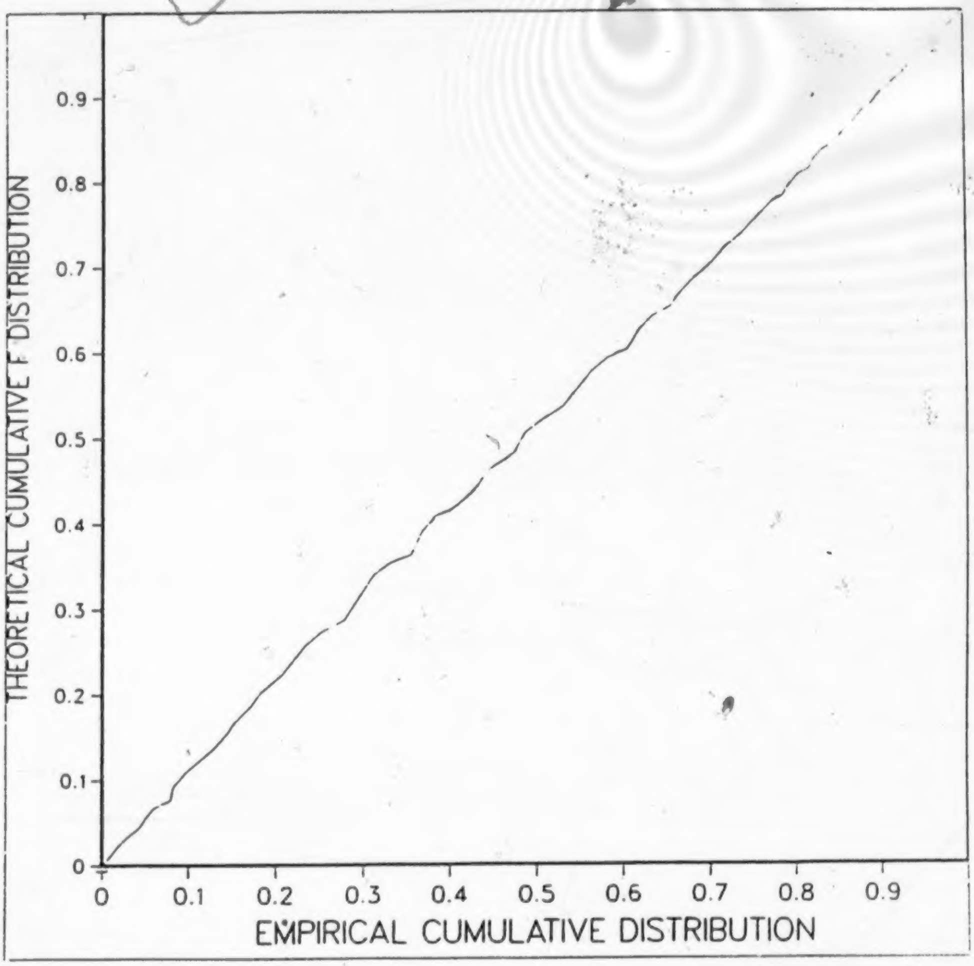

Figure A1.1 Overall goodness-of-fit X^2 testing no effects present (true model): $p = 0.1$, 40 observations per cell, 2×2×2 table (case 2.1).

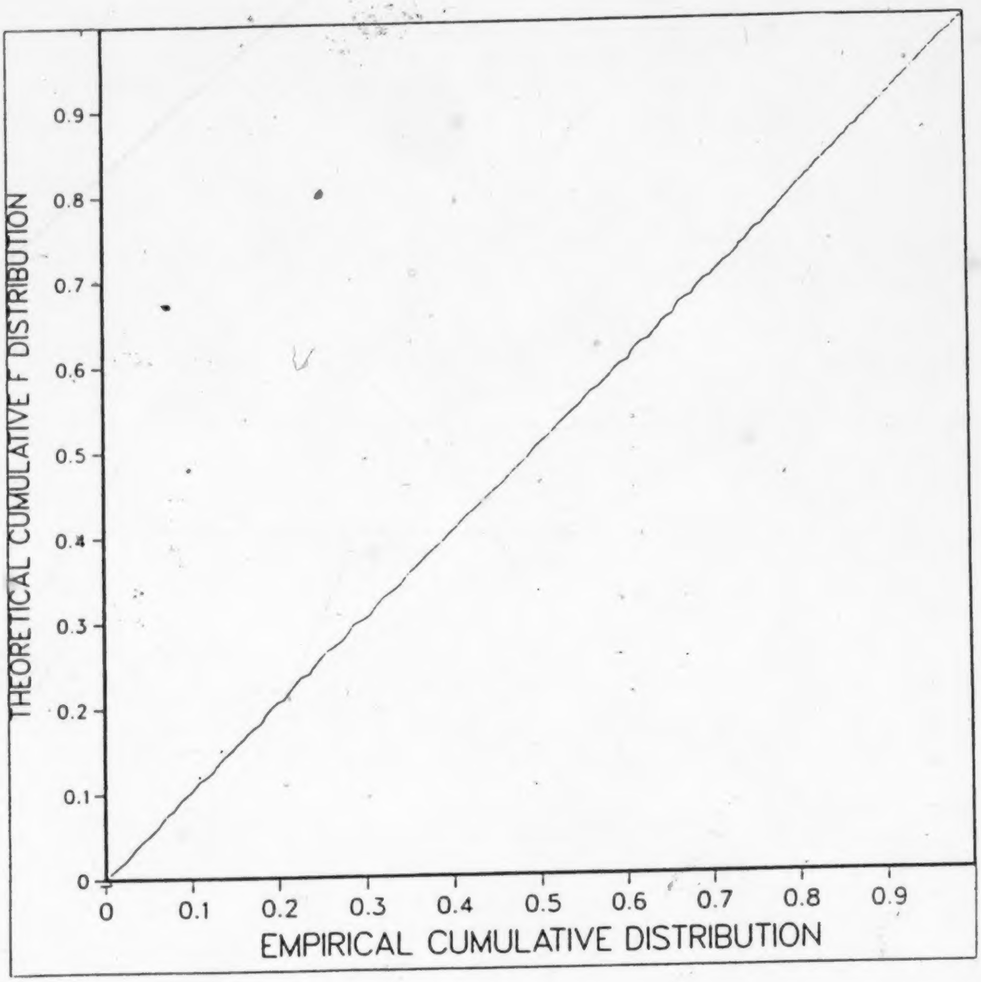

Figure A1.2 Overall goodness-of-fit X^2 testing no effects present (true model): $p = 0.25$, 40 observations per cell, 2×2×2 table (case 2.2).

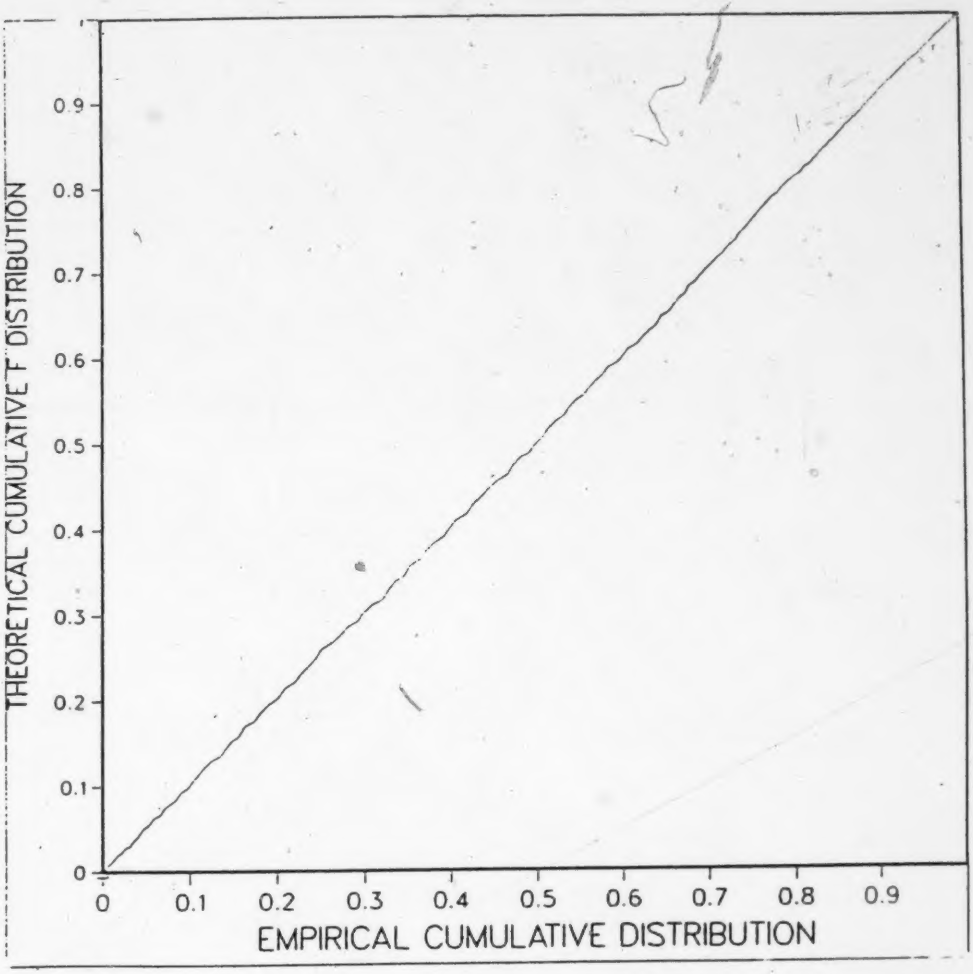

Figure A1.3 Overall goodness-of-fit X^2 testing no effects present (true model): $p = 0.5$, 40 observations per cell, 2×2×2 table (case 2.3).

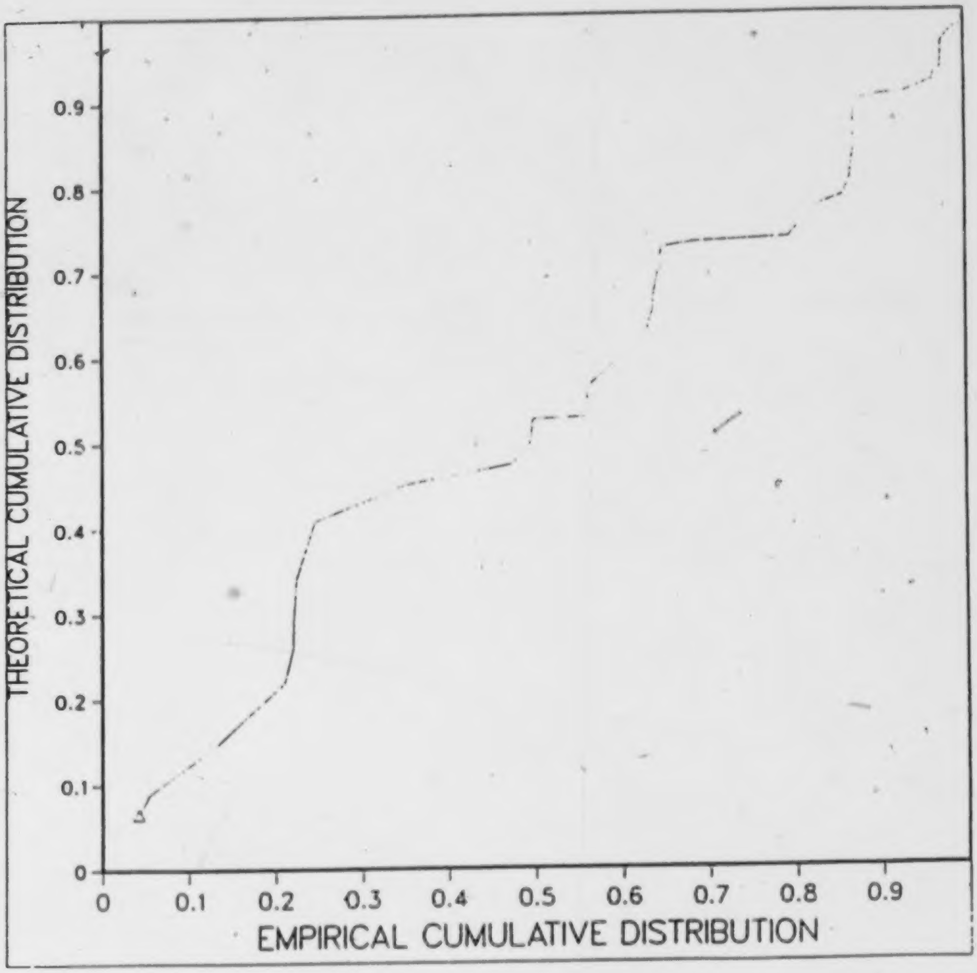

Figure A1.4 Overall goodness-of-fit X^2 testing no effects present (true model): $p = 0.1$, 10 observations per cell, 2×2 table (case 1.2).

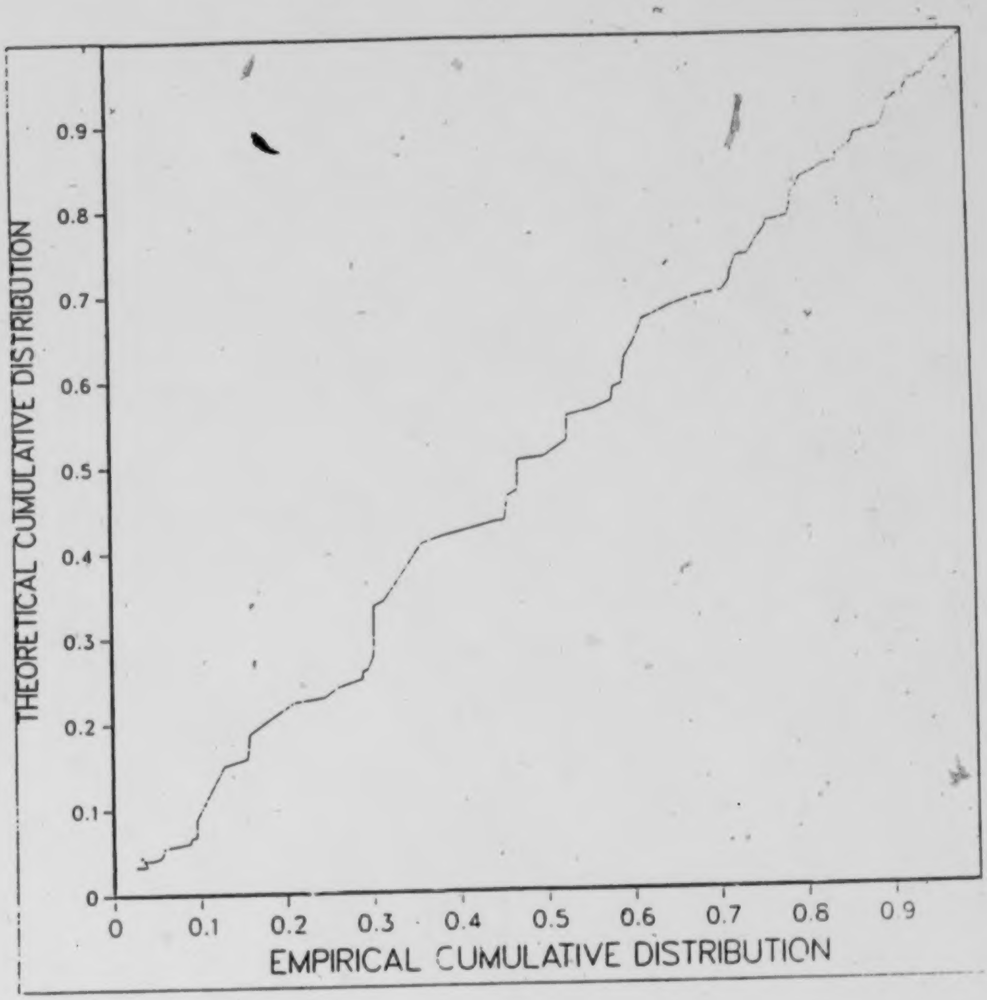

Figure A1.5 Overall goodness-of-fit X^2 testing no effects present (true model): $p = 0.25$, 10 observations per cell, 2×2 table (case 1.4).

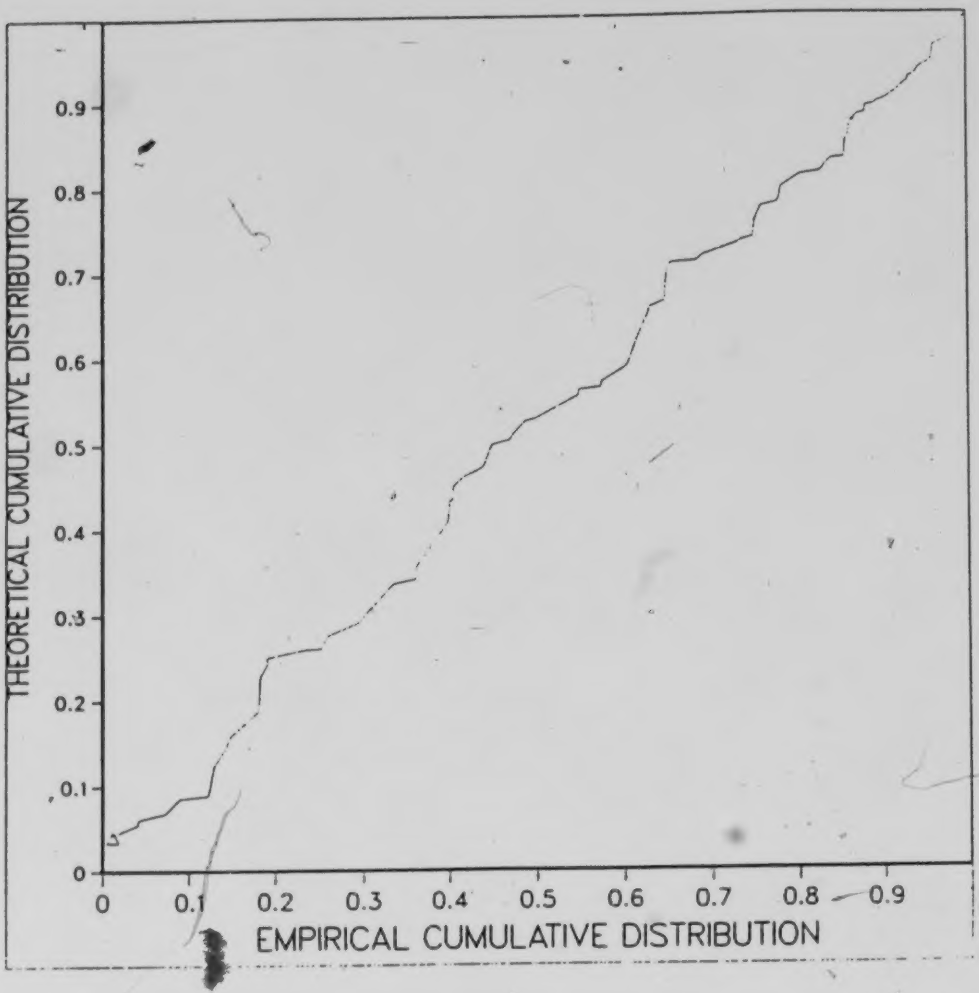

Figure A1.6 Overall goodness-of-fit X^2 testing no effects present (true model): $p = 0.5$, 10 observations per cell, 2×2 table (case 1.6).

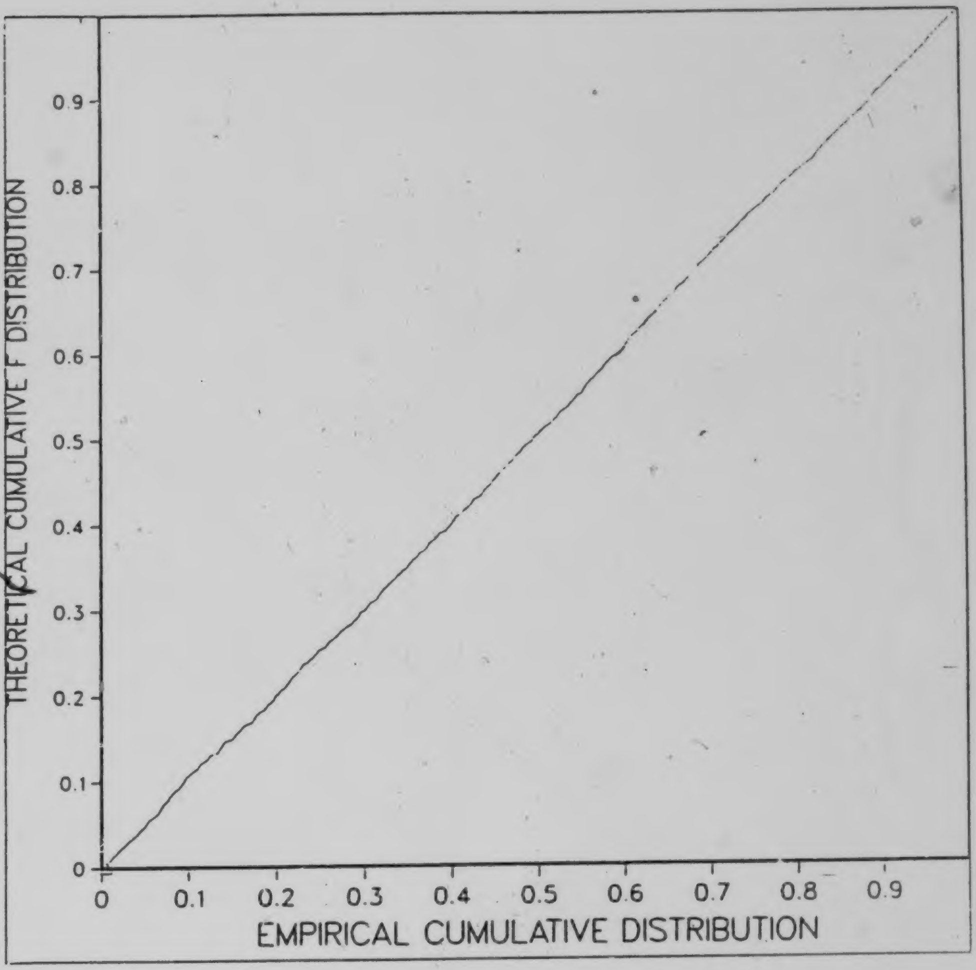

Figure A1.7 ANOVA F testing for row effects with no effects present: $p = 0.1$, 40 observations per cell, 2×2×2 table (case 2.1).

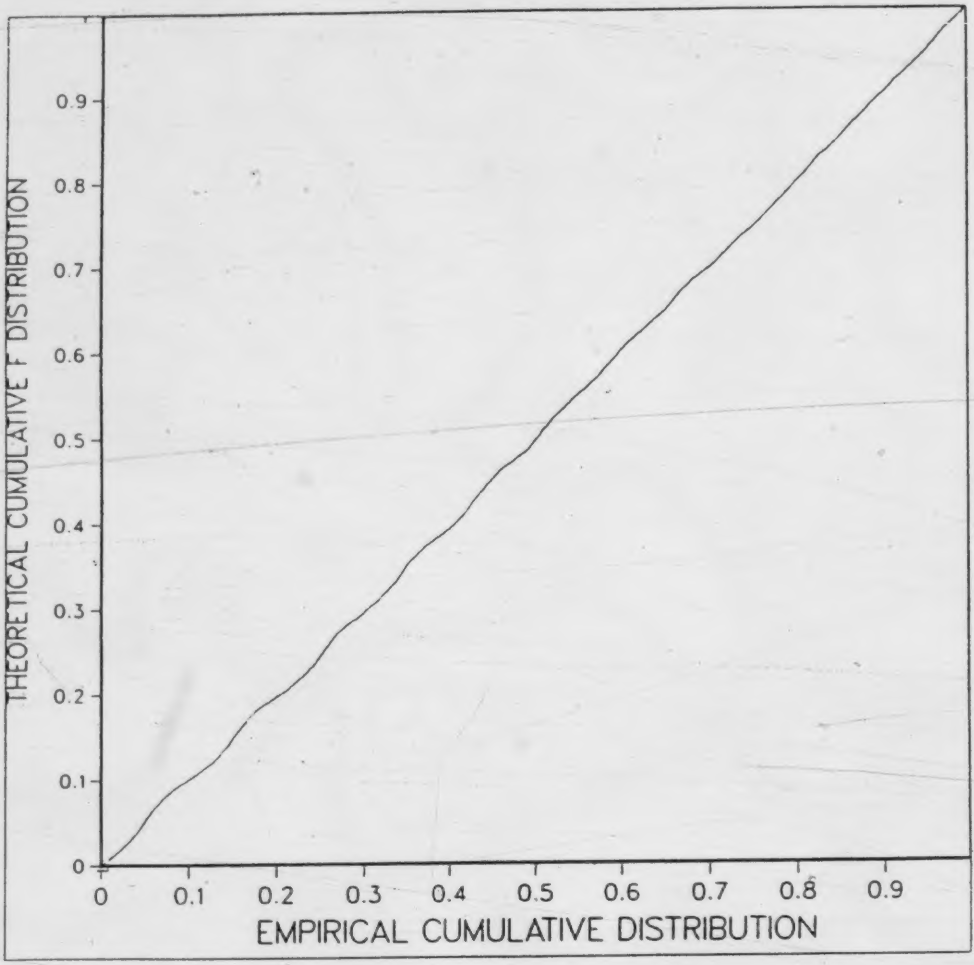

Figure A1.8 ANOVA F testing for row effects with no effects present: $p = 0.25$, 40 observations per cell, 2×2×2 table (case 2.2).

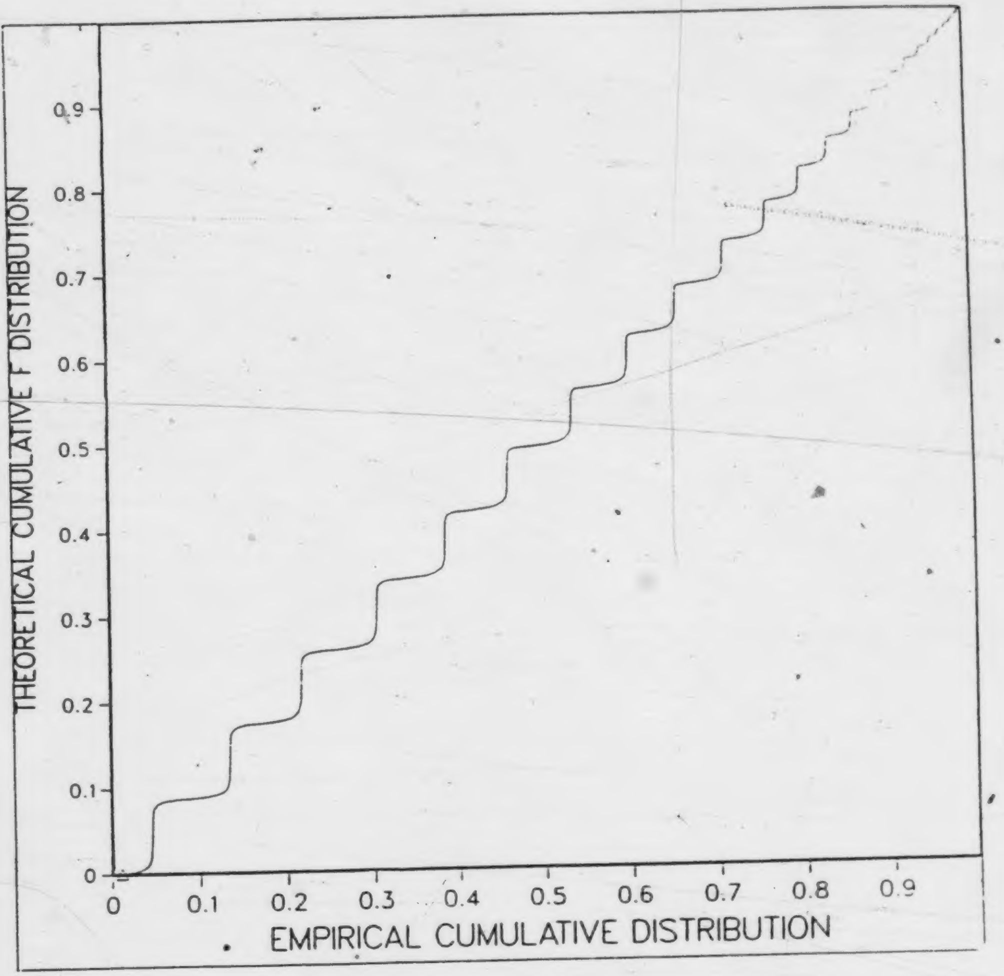

Figure A1.9 ANOVA F testing for row effects with no effects present: $p = 0.5$, 40 observations per cell, 2×2×2 table (case 2.3).

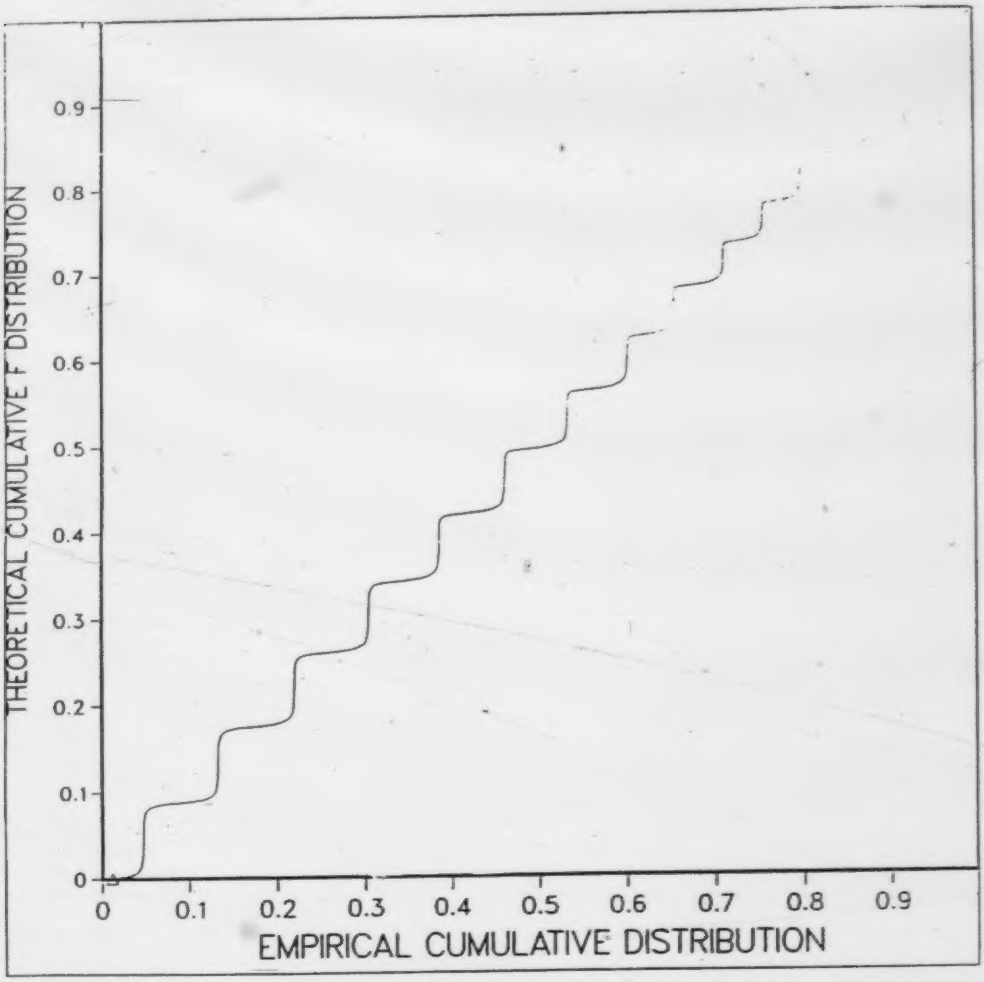

Figure A1.10 ANOVA F testing with arcsine transformation for row-column interaction effects with no effects present: $p = 0.5$, 40 observations per cell, 2×2×2 table (case 2.3).

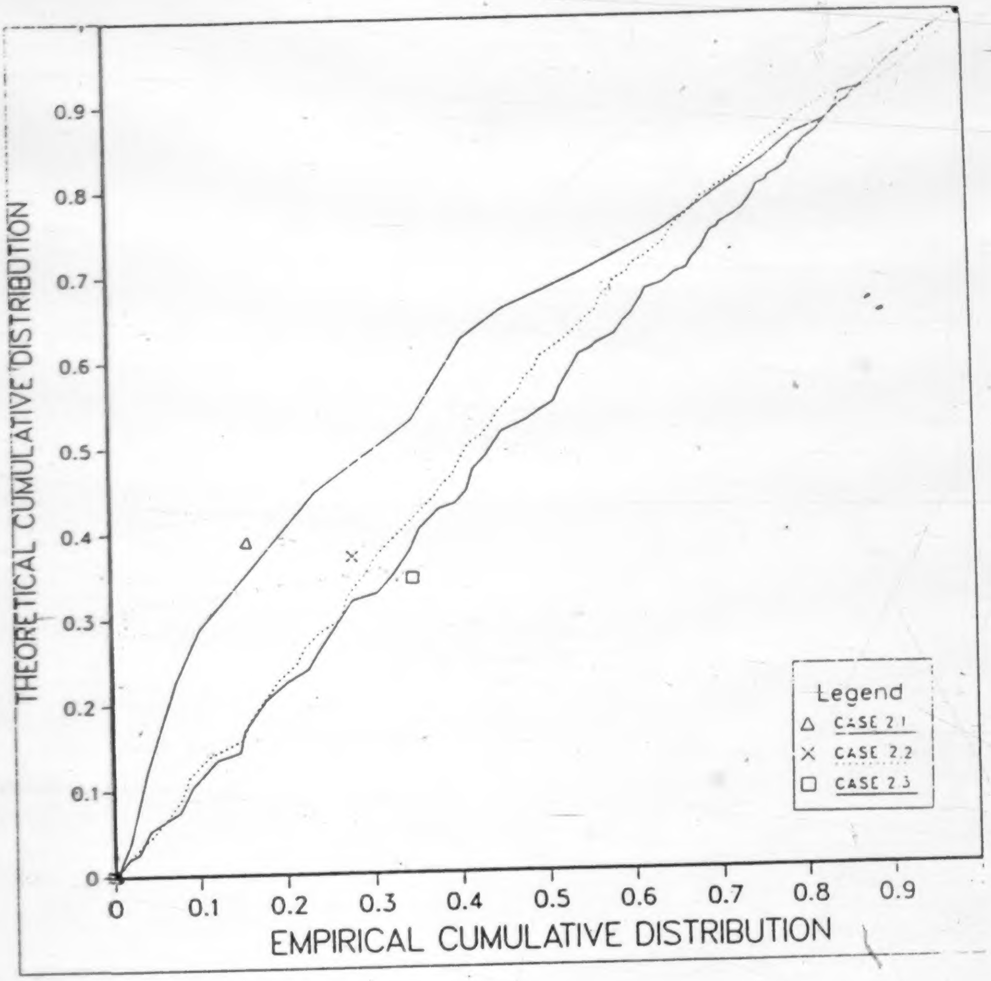

Figure A1.11 G^2 overall goodness-of-fit: 2×2×2 table, 10 observations per cell, $p = 0.1$, $p = 0.25$, $p = 0.5$ (cases 2.1, 2.2, 2.3, respectively).

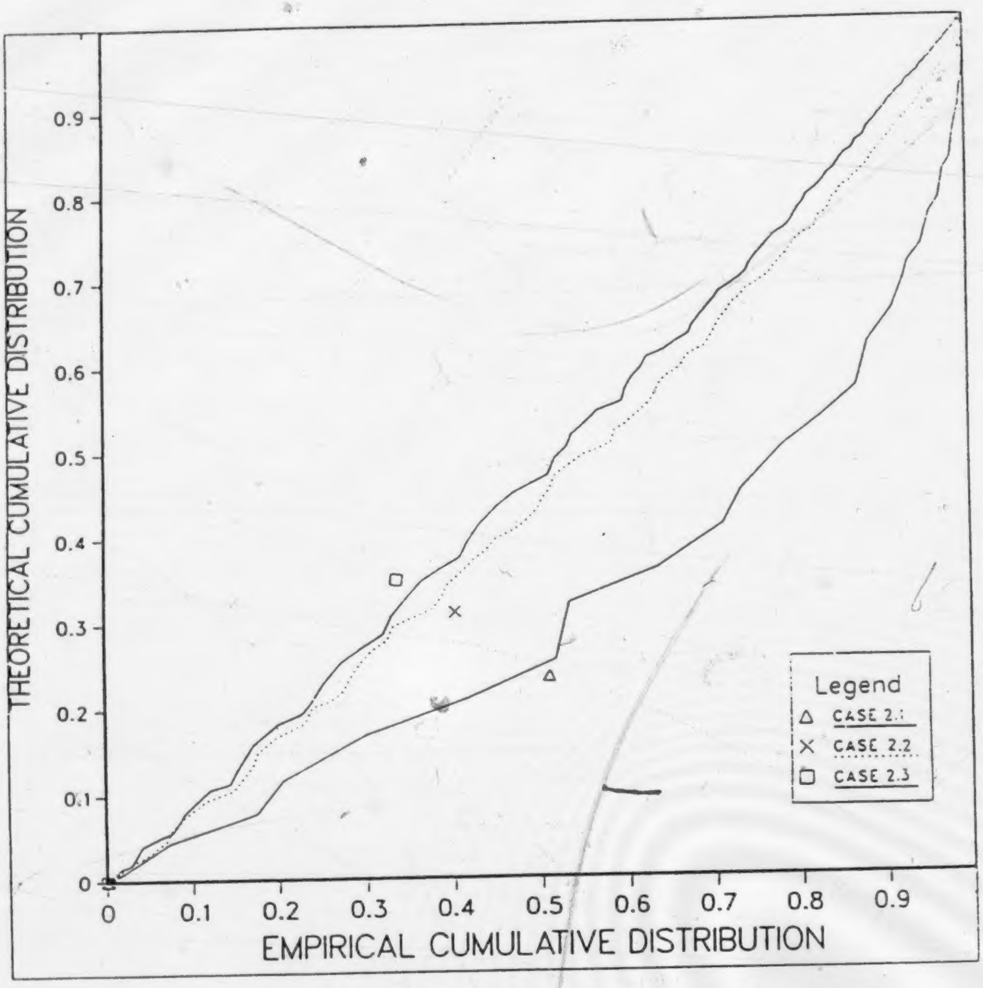

Figure A1.12 G^2 overall goodness-of-fit after addition of 0.5 to each cell total: 2×2×2 table, 10 observations per cell, $p = 0.1$, $p = 0.25$, $p = 0.5$ (cases 2.1, 2.2, 2.3, respectively).

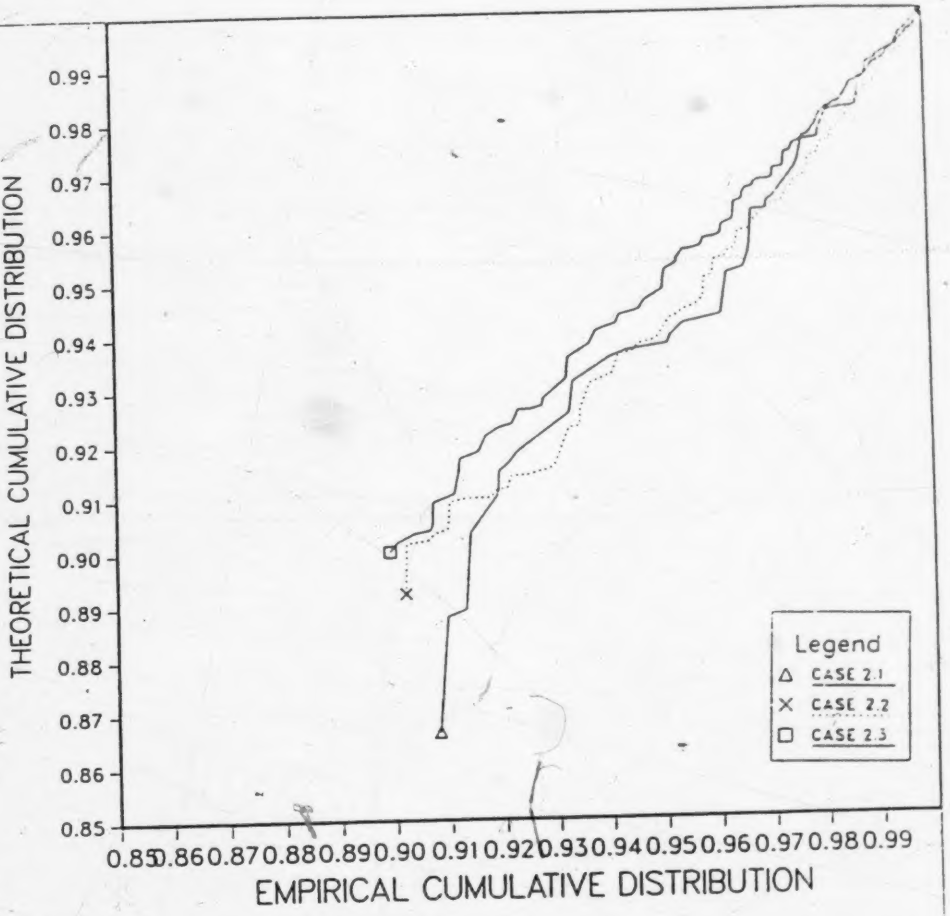

Figure A1.13 X^2 overall goodness-of-fit: 2×2×2 table, 10 observations per cell, $p = 0.1$, $p = 0.25$, $p = 0.5$ (cases 2.1, 2.2, 2.3, respectively).

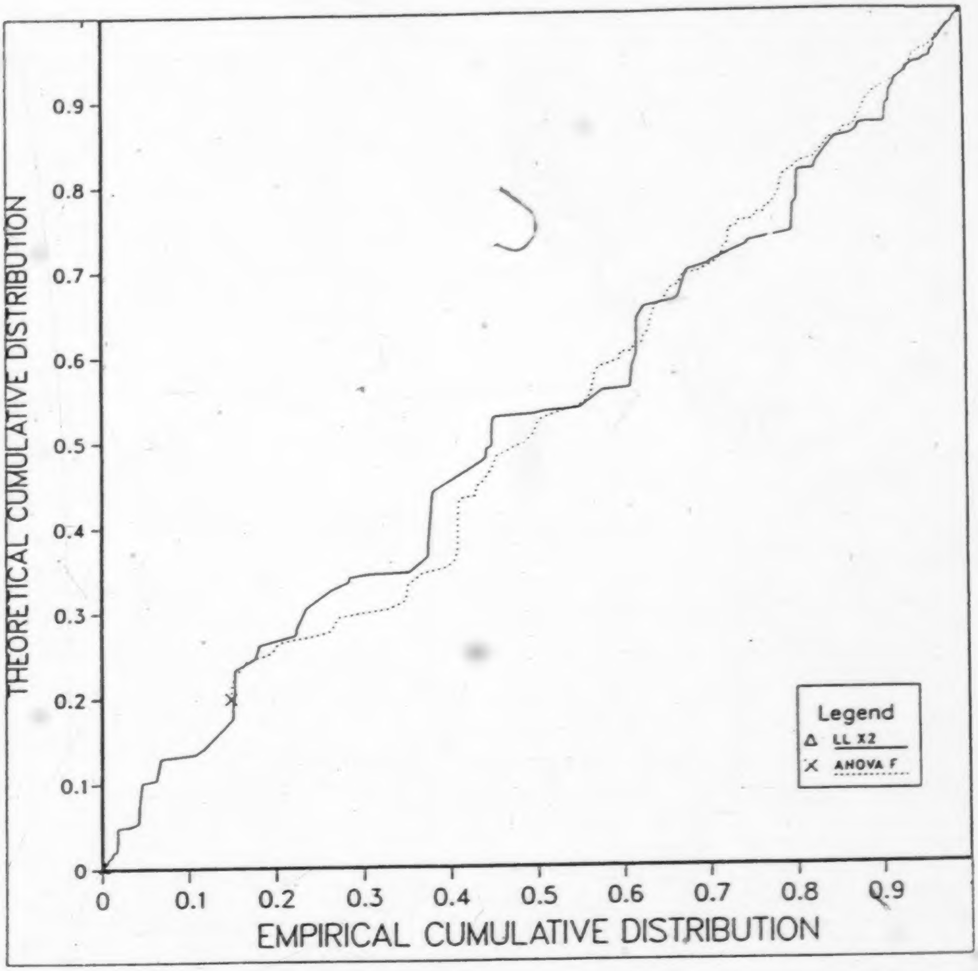

Figure A1.14 X^2 overall goodness-of-fit and ANOVA F for row effects: 2×2×2 table, 10 observations per cell, $p = 0.1$ (case 2.1).

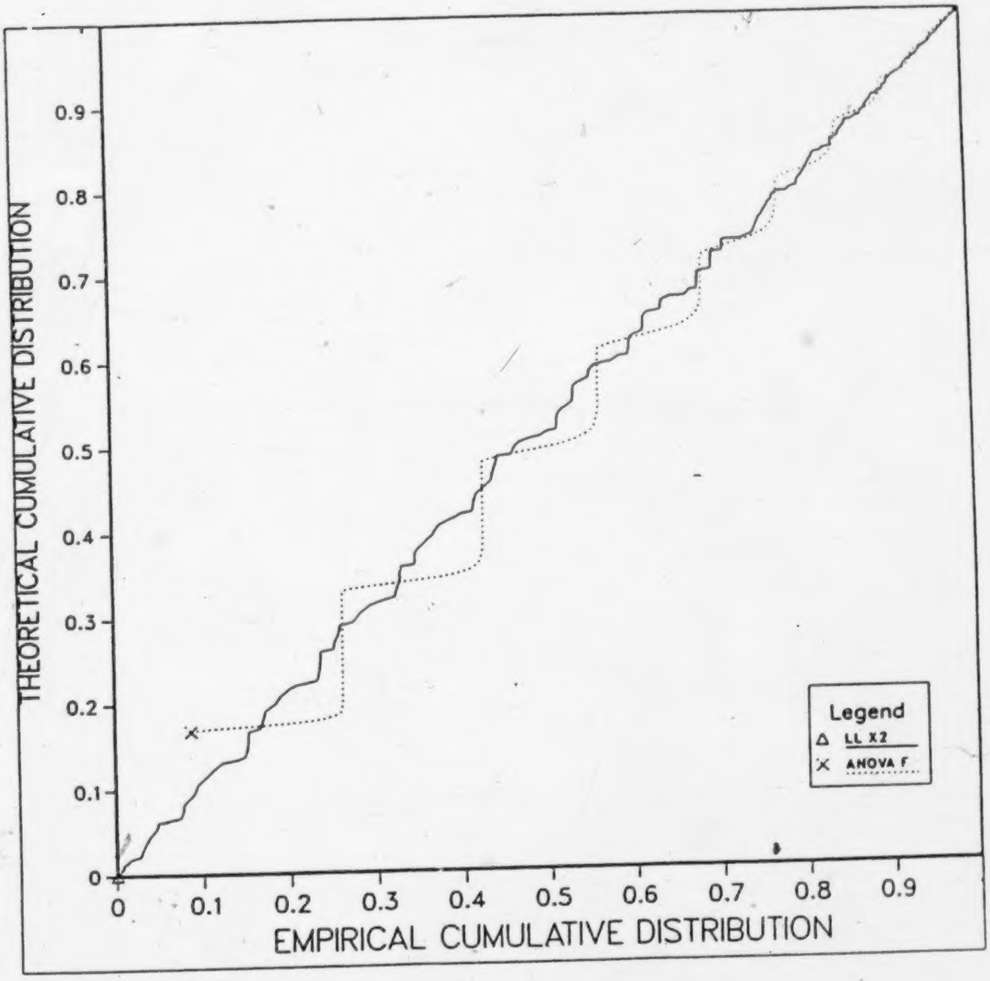

Figure A1.15 X^2 overall goodness-of-fit and ANOVA F for row effects: 2×2×2 table, 10 observations per cell, $p = 0.5$ (case 2.3).

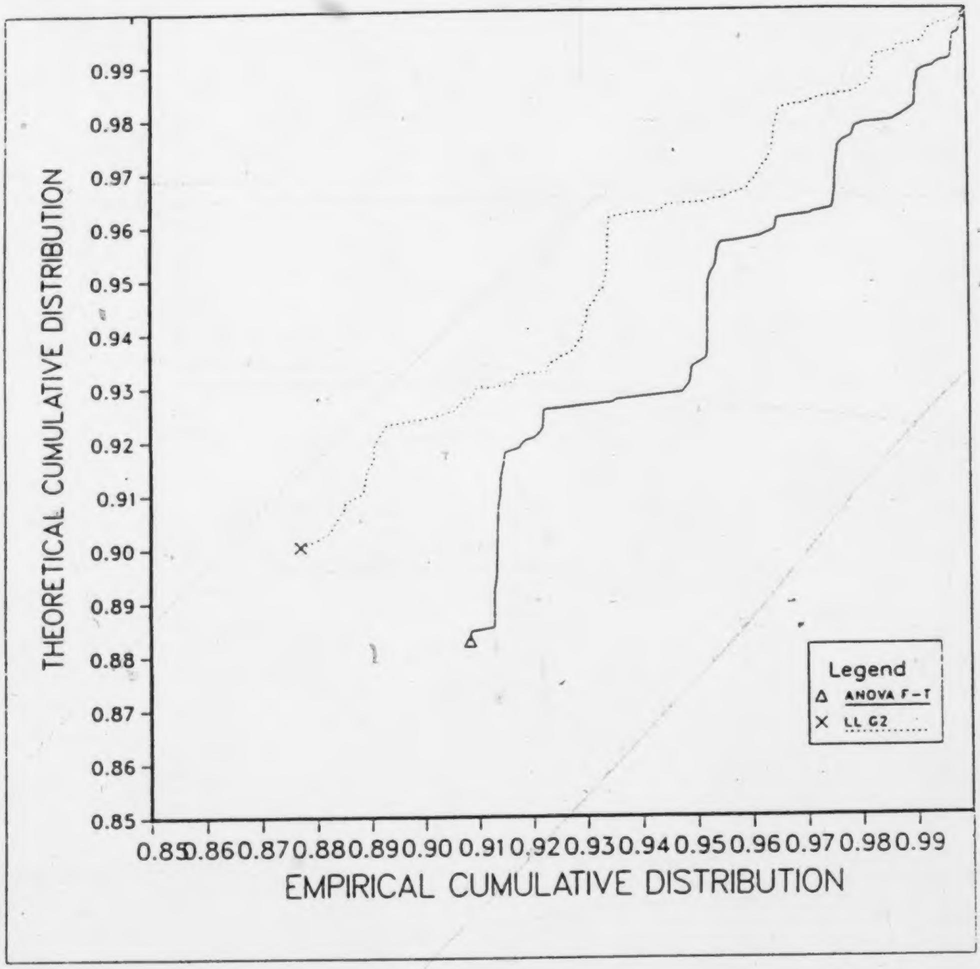

Figure A1.16 Conditional G^2 backward and ANOVA F with arcsine transformation testing for row effects: 2×2×2 table, 10 observations per cell, $p = 0.1$ (case 2.1).

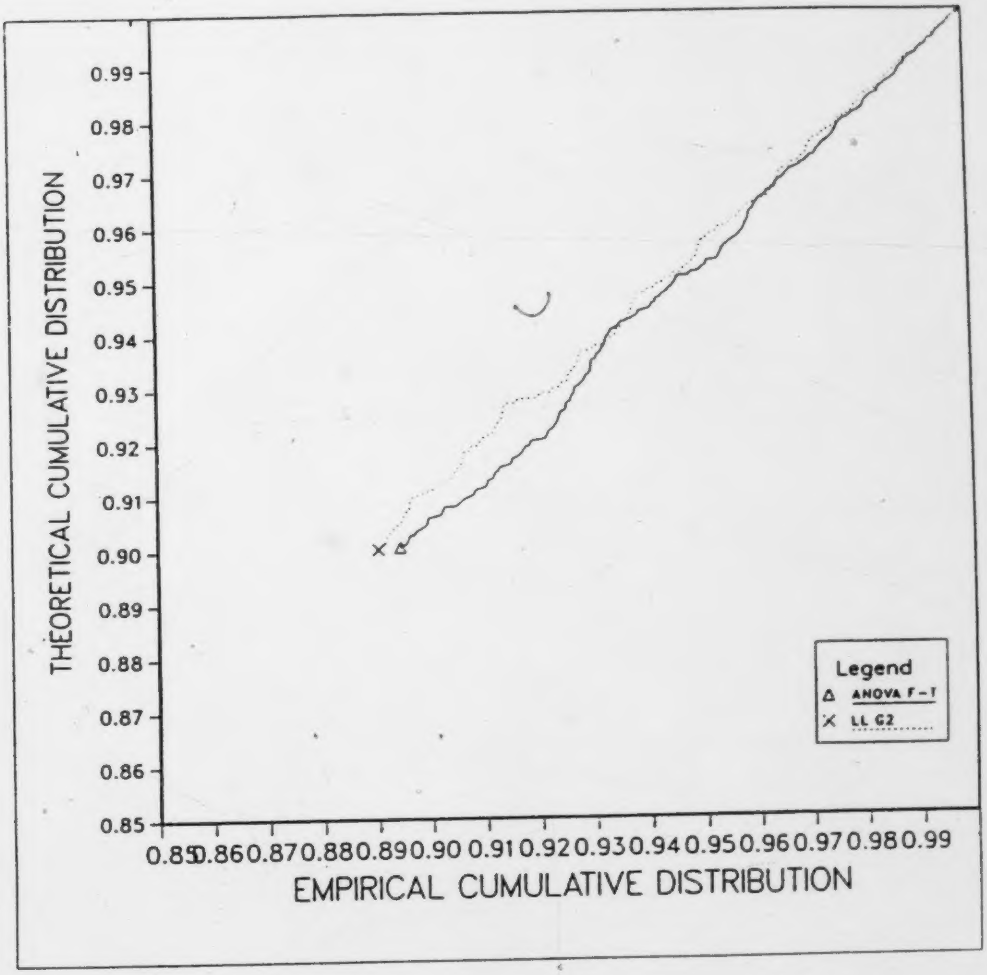

Figure A1.17 Conditional G^2 backward and ANOVA F with arcsine transformation testing for row effects: 2×2×2 table, 10 observations per cell, $p = 0.25$ (case 2.2).

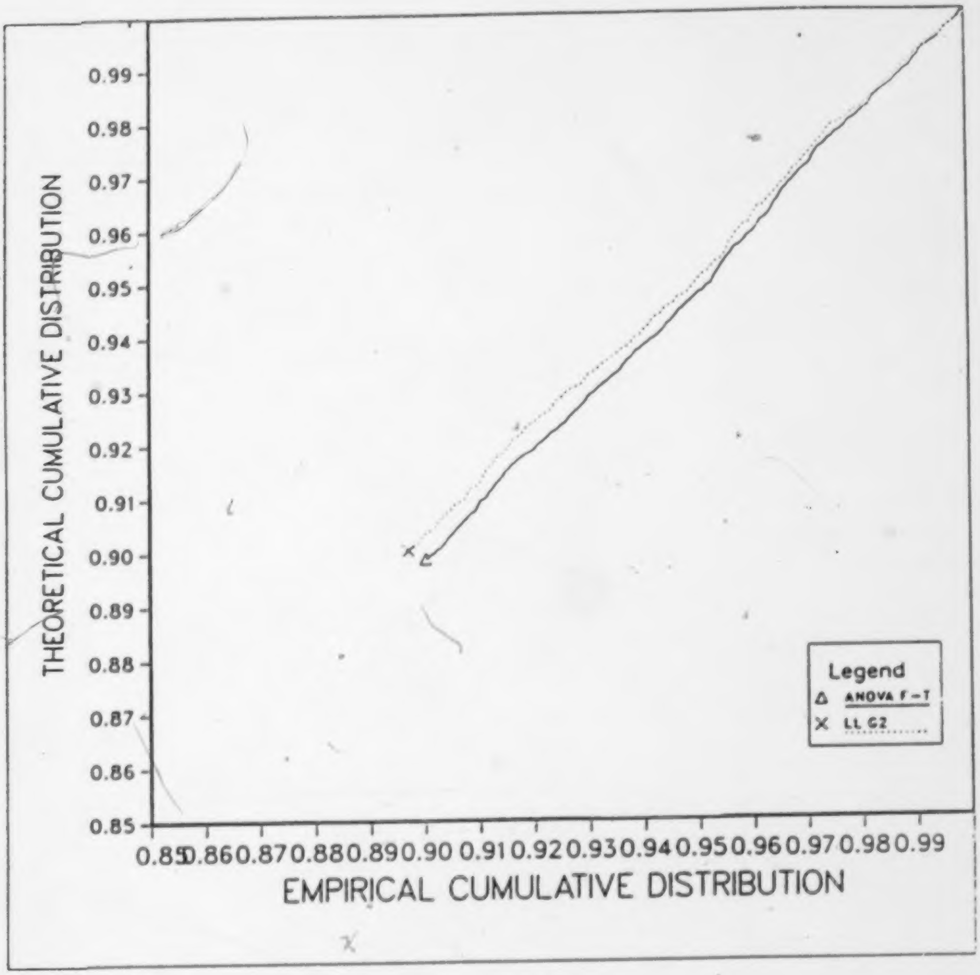

Figure A1.18 Conditional G^2 backward and ANOVA F with arcsine transformation testing for row effects: 2×2×2 table, 40 observations per cell, $p = 0.25$ (case 2.2).

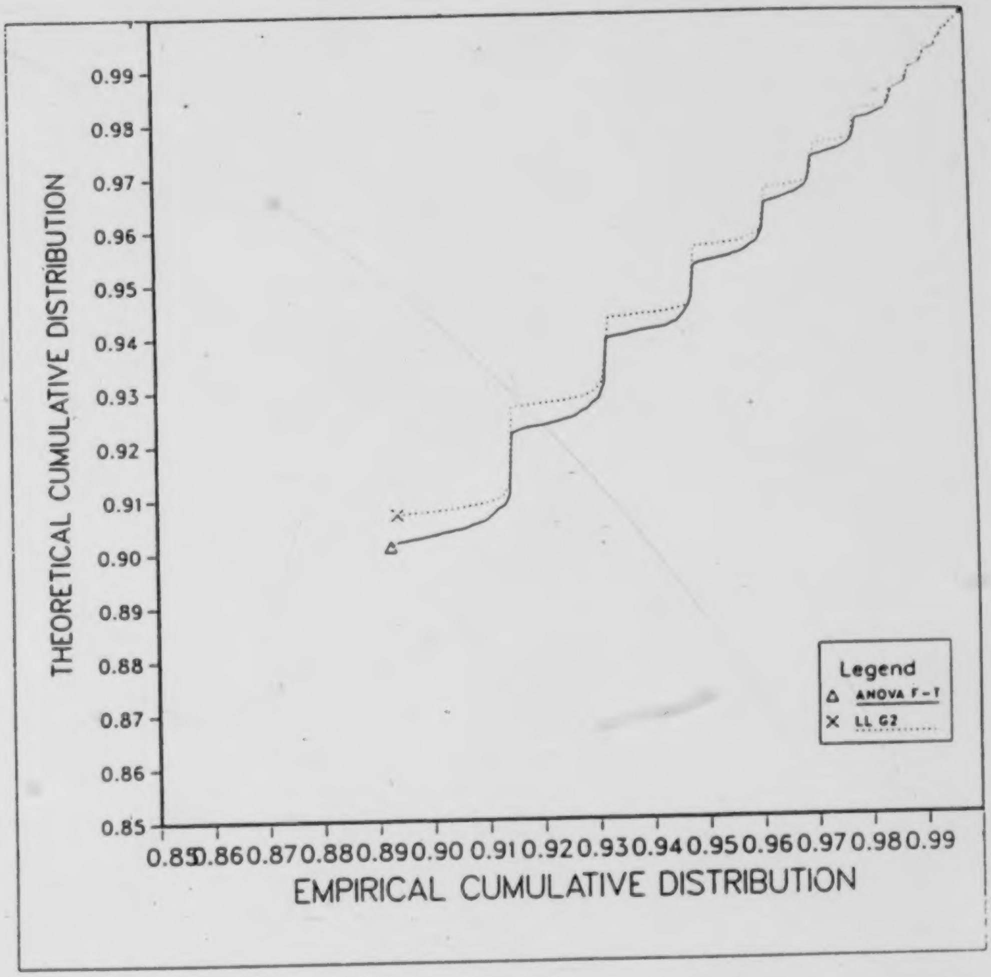

Figure A1.19 Conditional G^2 backward and ANOVA F with arcsine transformation testing for row effects: 2×2×2 table, 40 observations per cell, $p = 0.5$ (case 2.3).

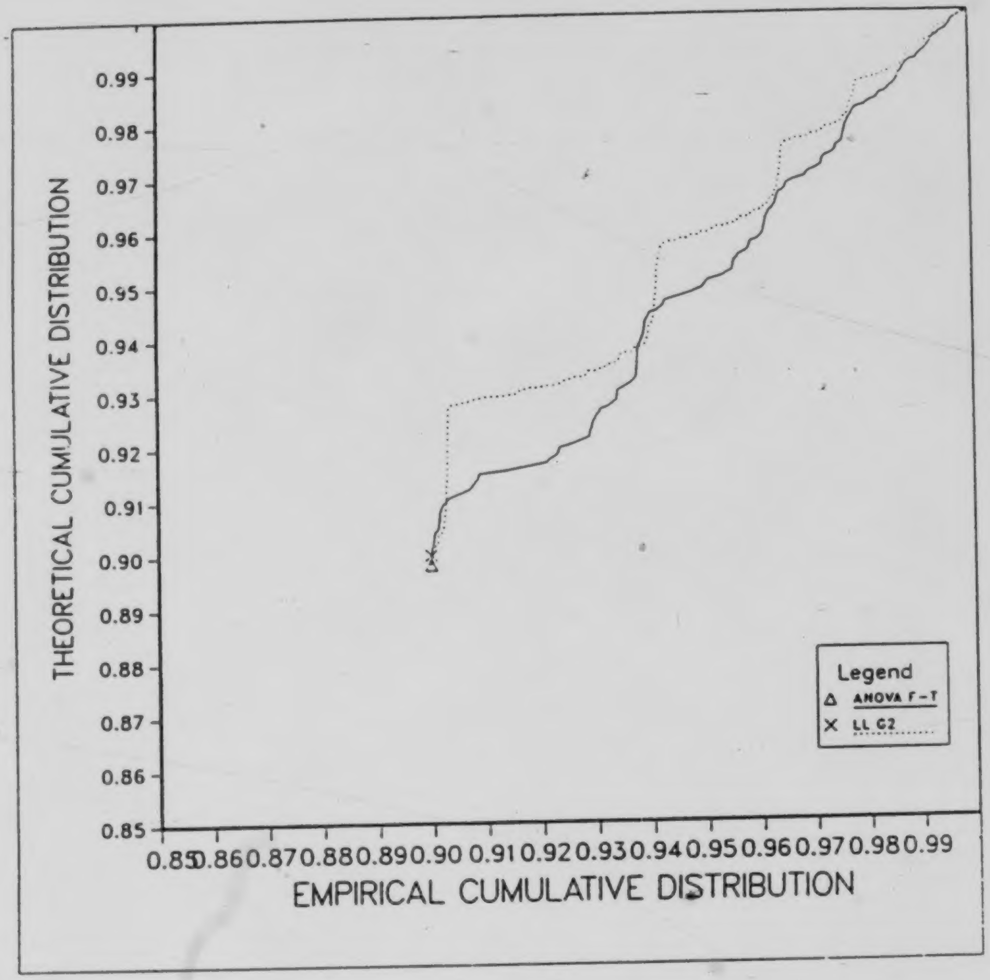

Figure A1.20 Conditional G^2 backward and ANOVA F with arcsine transformation testing for row effects: 2×2×2 table, 40 observations per cell, $p = 0.5$ (case 2.3).

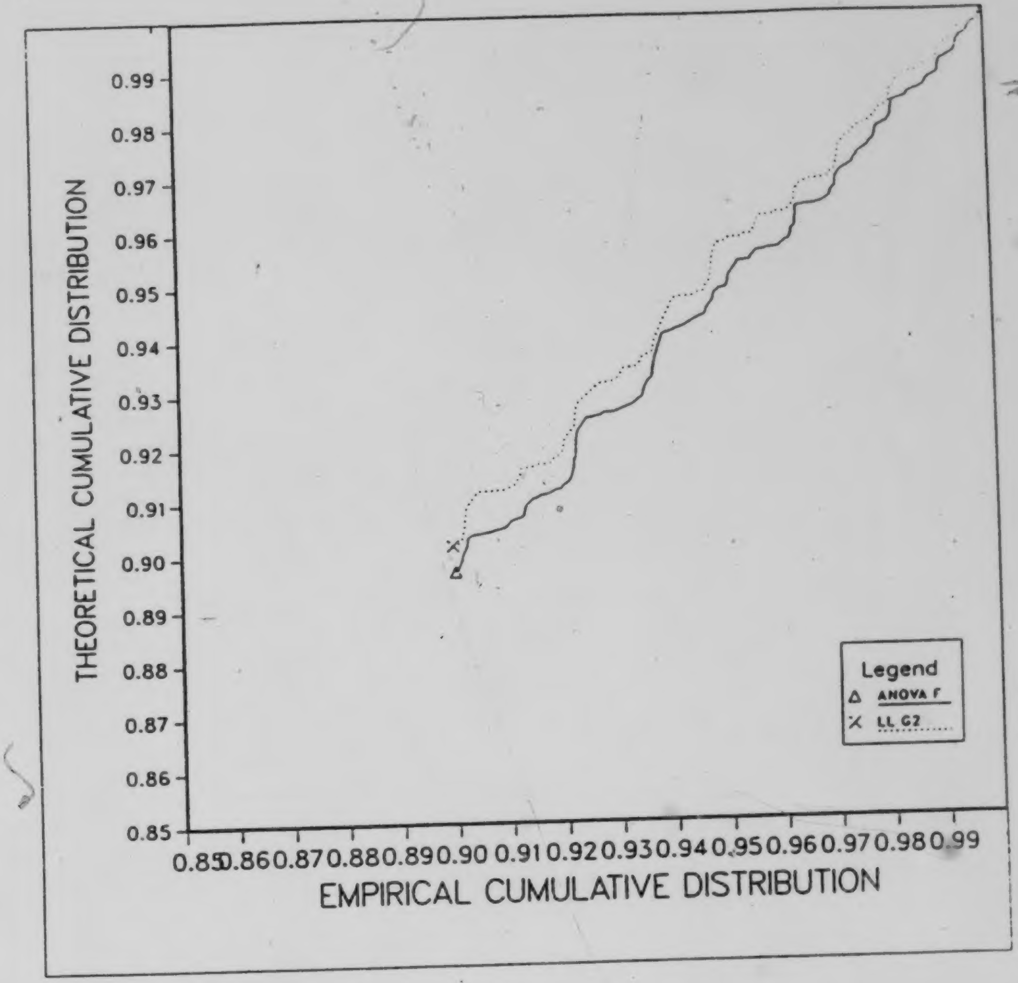

Figure A1.21 Conditional G^2 backward and ANOVA F for row effects: 2×2 table, 40 observations per cell, $p = 0.1$ (case 1.2).

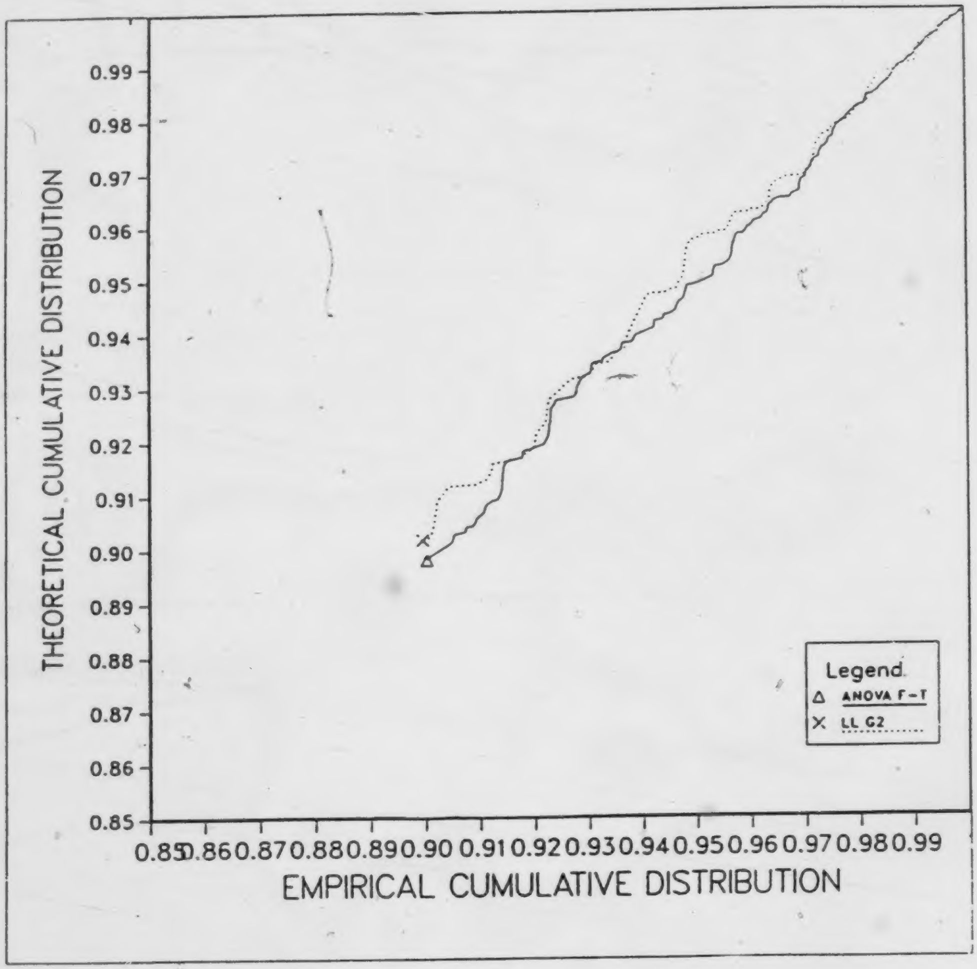

Figure A1.22 Conditional G^2 backward and ANOVA F with arcsine transformation for row effects: 2×2 table, 40 observations per cell, $p = 0.1$ (case 1.2).

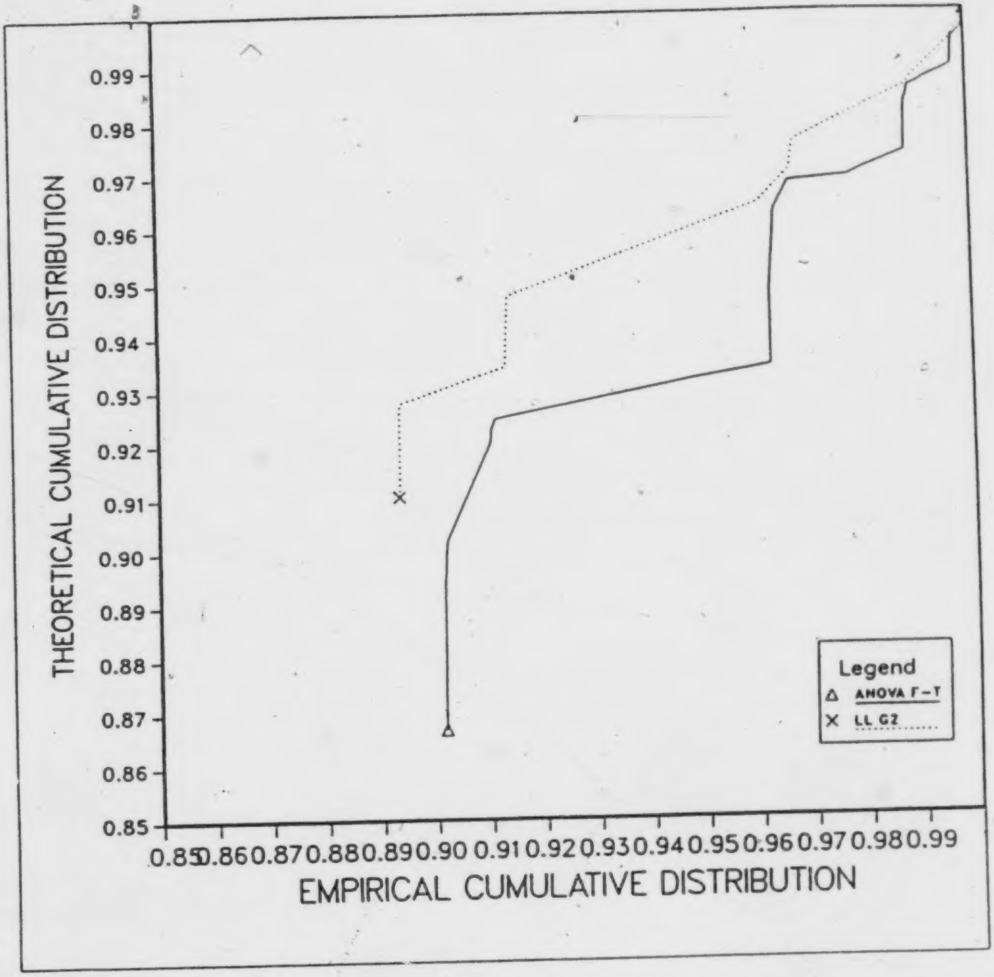

Figure A1.23 Conditional G^2 forward and ANOVA F with arcsine transformation testing for row effects: 2×2 table, 10 observations per cell, $p = 0.1$ (case 1.2).

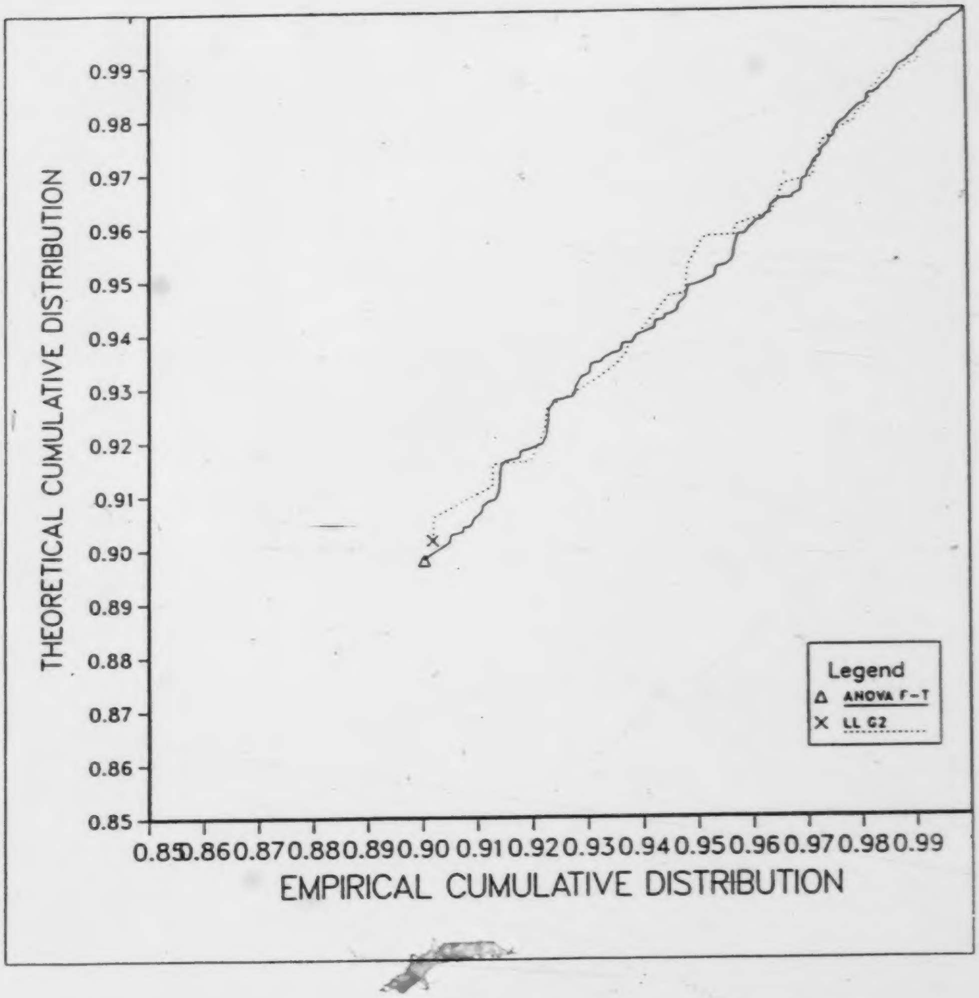

Figure A1.24 Conditional G^2 forward and ANOVA F with arcsine transformation testing for row effects: 2×2 table, 40 observations per cell, $p = 0.1$ (case 1.2).

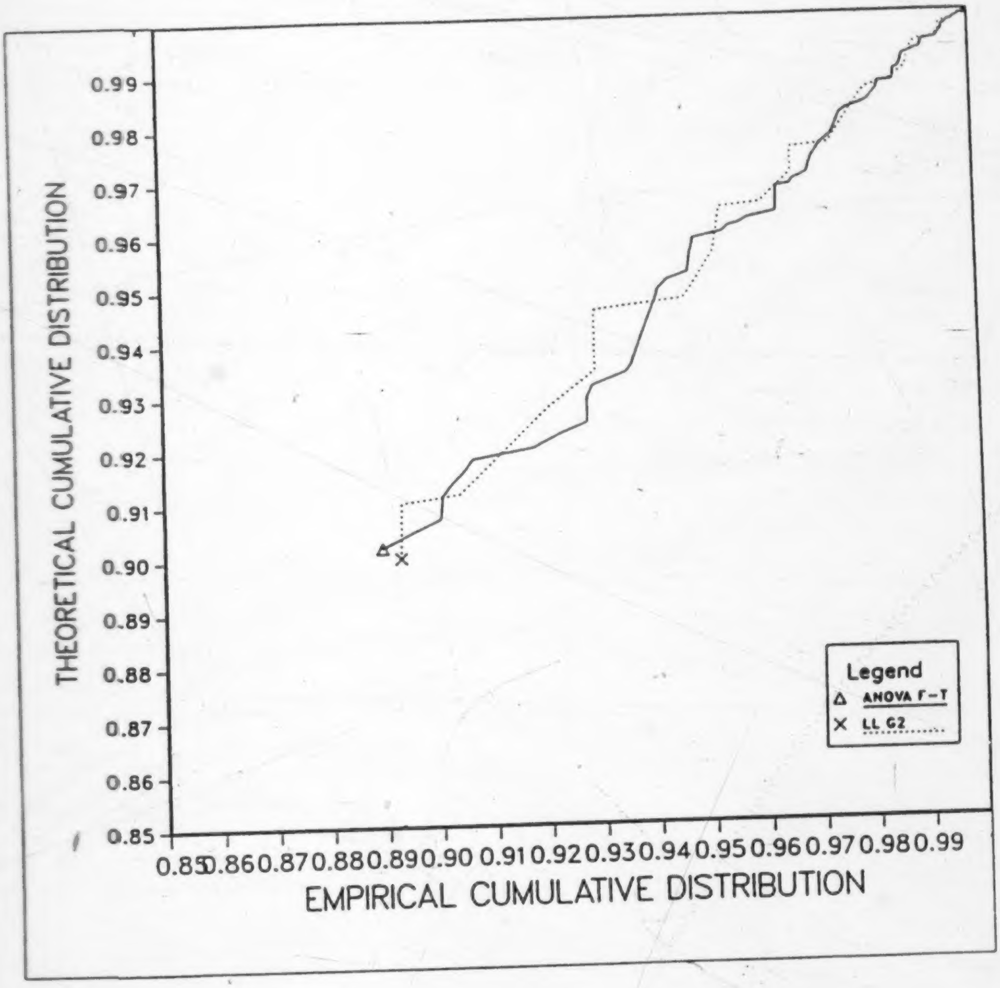

Figure A1.25 Conditional G^2 forward and ANOVA F with arcsine transformation testing for row effects: 2×2 table, 10 observations per cell, $p = 0.25$ (case 1.4).

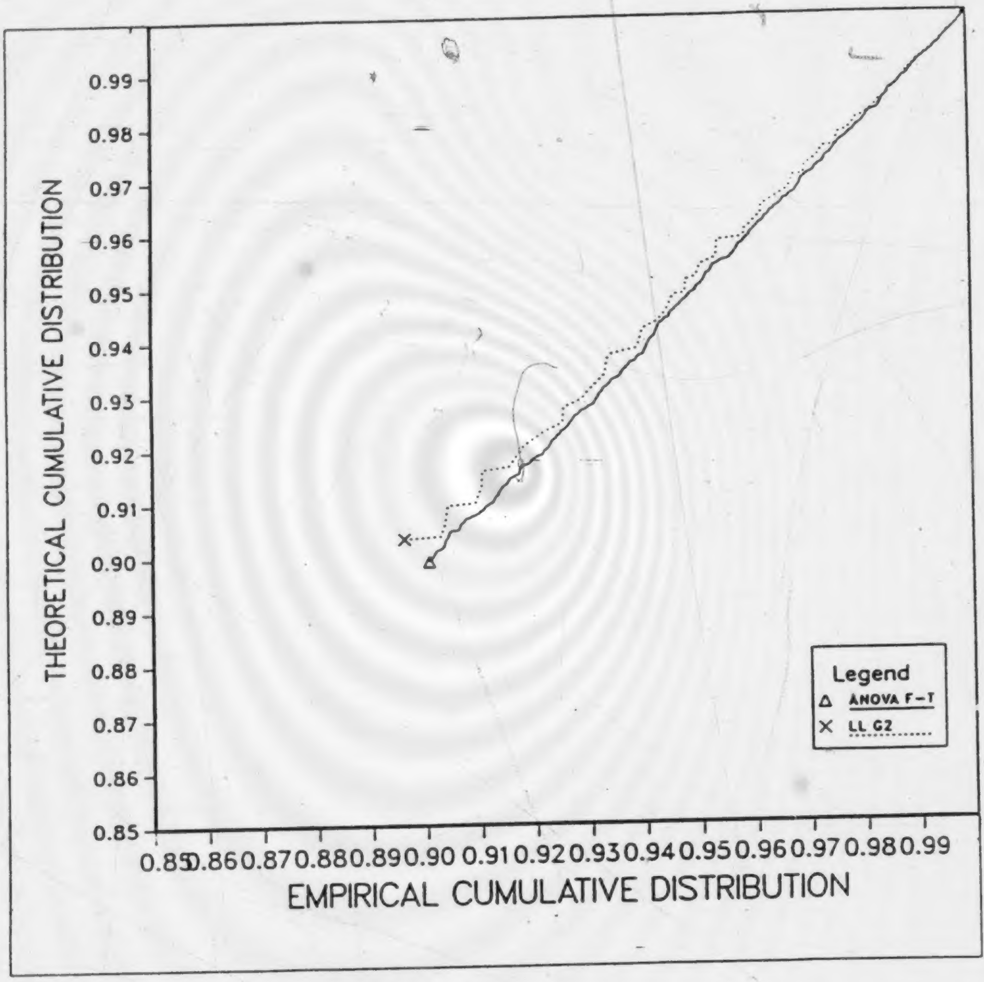

Figure A1.26 Conditional G^2 forward and ANOVA F with arcsine transformation testing for row effects: 2×2 table, 40 observations per cell, $p = 0.25$ (case 1.4).

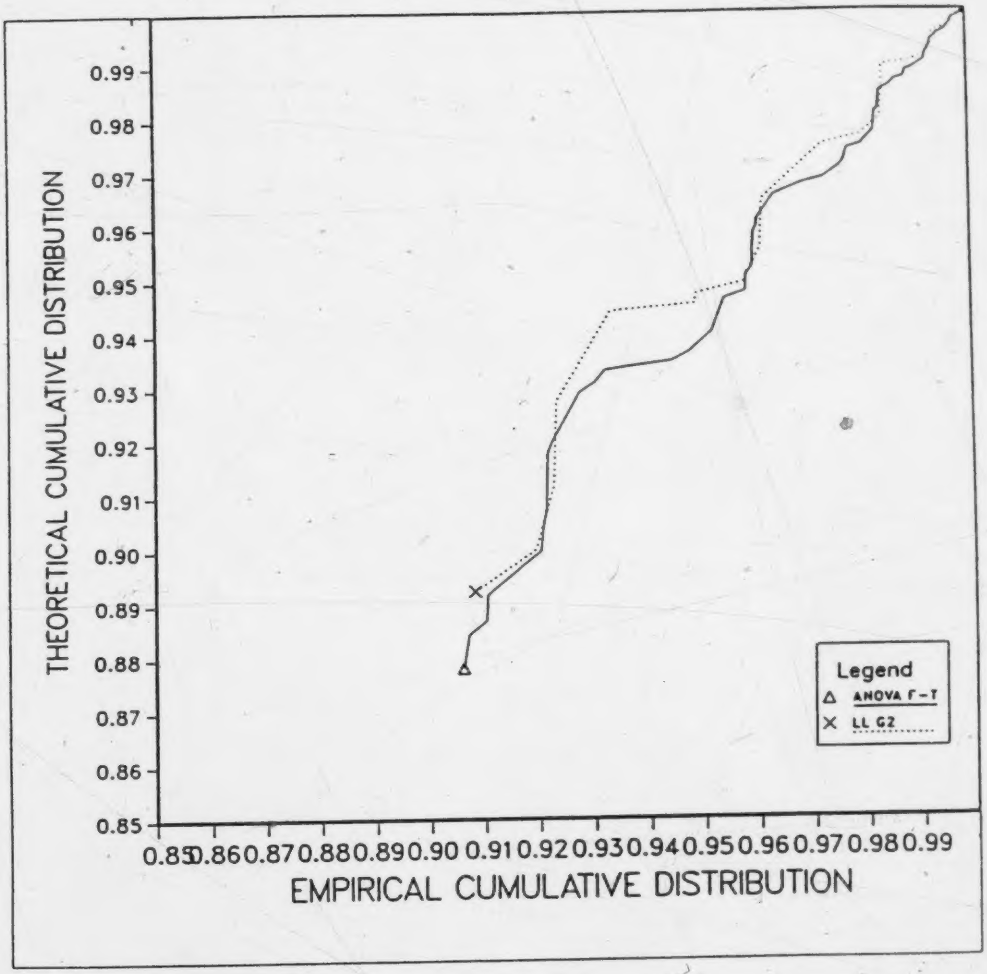

Figure A1.27 Conditional G^2 forward and ANOVA F with arcsine transformation testing for row effects: 2×2 table, 10 observations per cell, $p = 0.5$ (case 1.6).

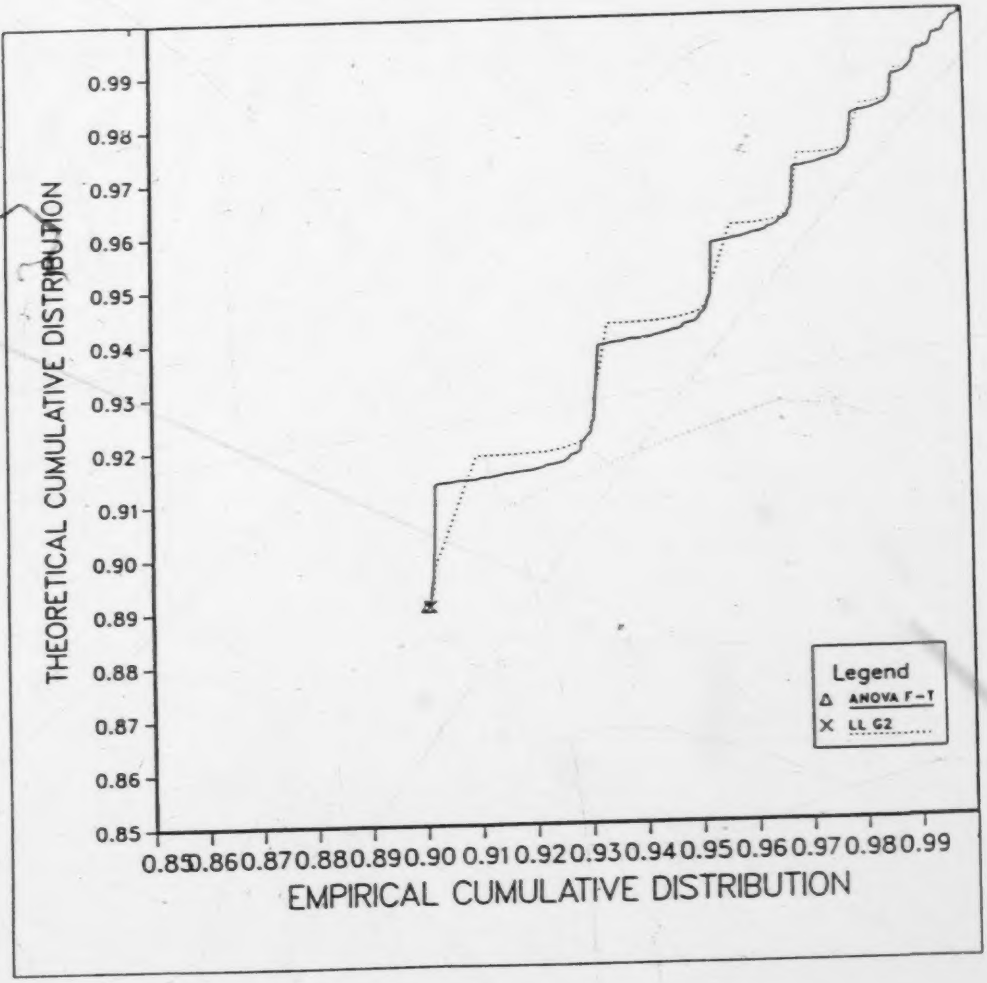

Figure A1.28 Conditional G^2 forward and ANOVA F with arcsine transformation testing for row effects: 2×2 table, 40 observations per cell, $p = 0.5$ (case 1.6).

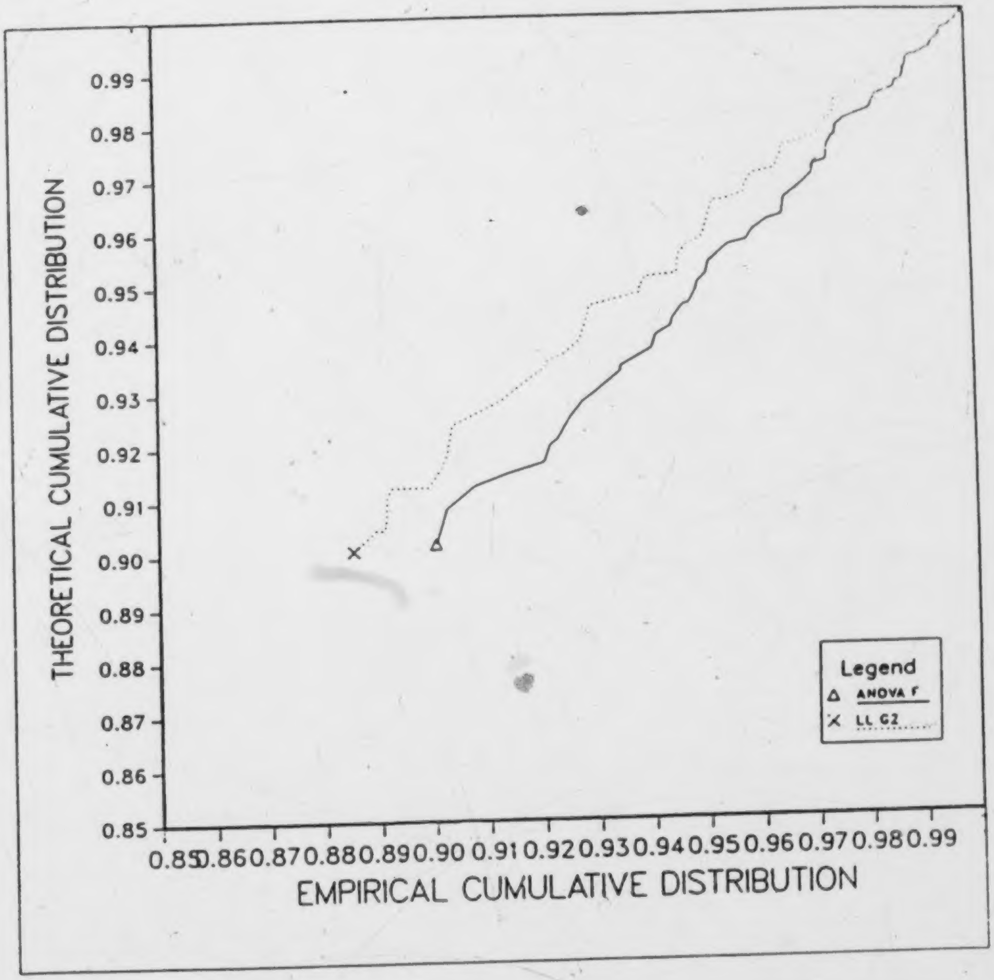

Figure A1.29 Conditional G^2 backward and ANOVA F for row effects:
2×2 table, 10 observations per cell, $p = 0.25$ (case 1.4).

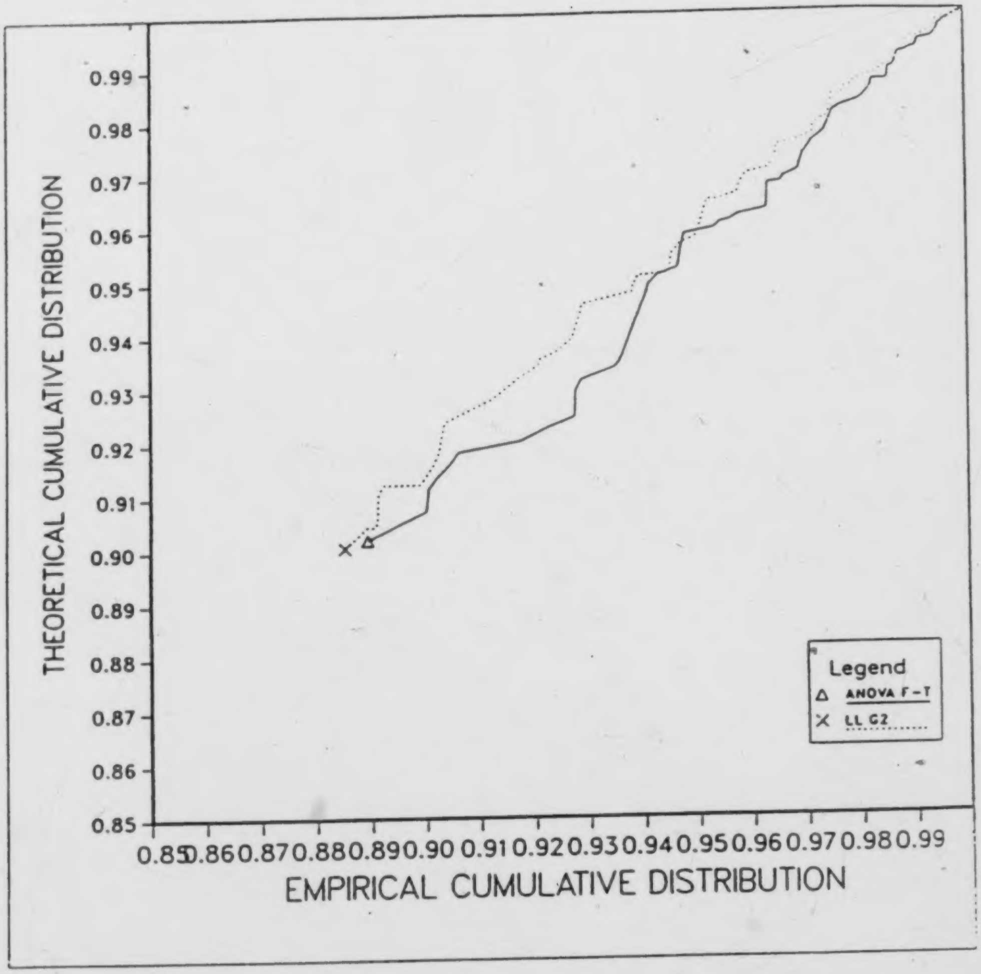

Figure A1.30 Conditional G^2 backward and ANOVA F with arcsine transformation testing for row effects: 2×2 table, 10 observations per cell, $p = 0.25$ (case 1.4).

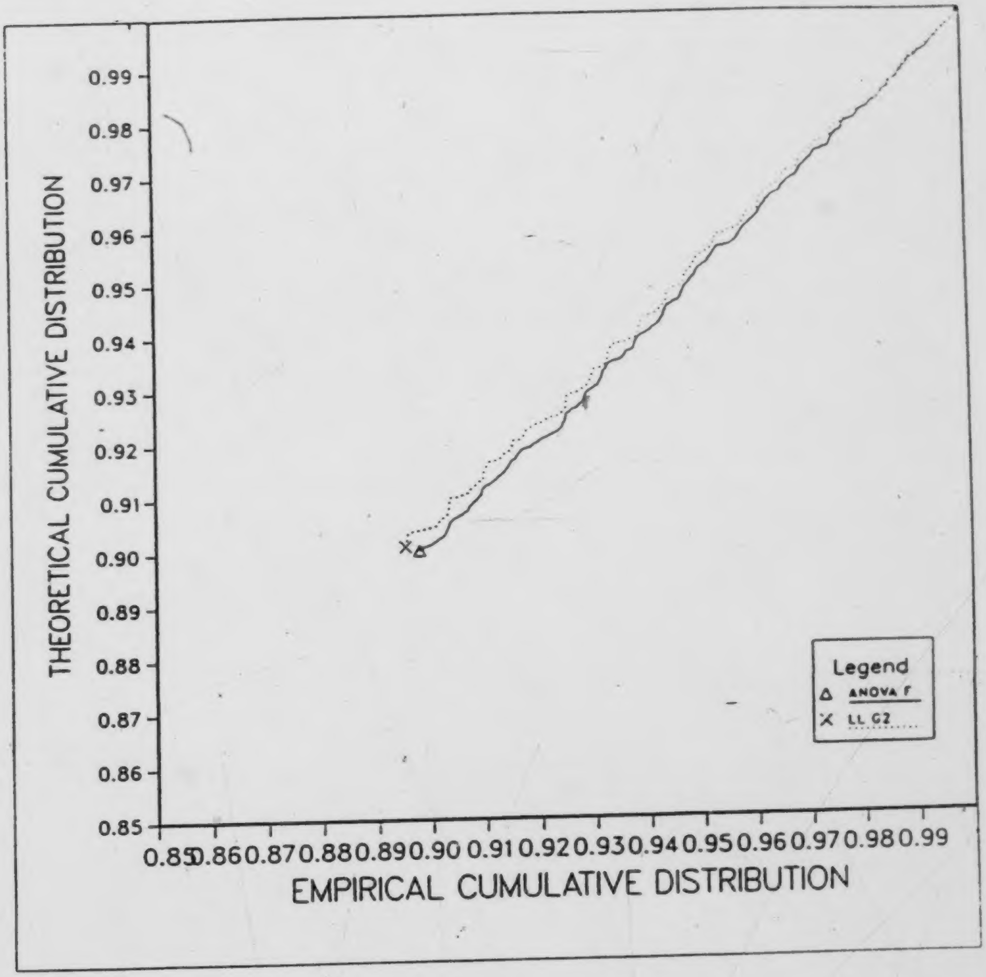

Figure A1.31 Conditional G^2 backward and ANOVA F for row effects: 2×2 table, 40 observations per cell, $p = 0.25$ (case 1.4).

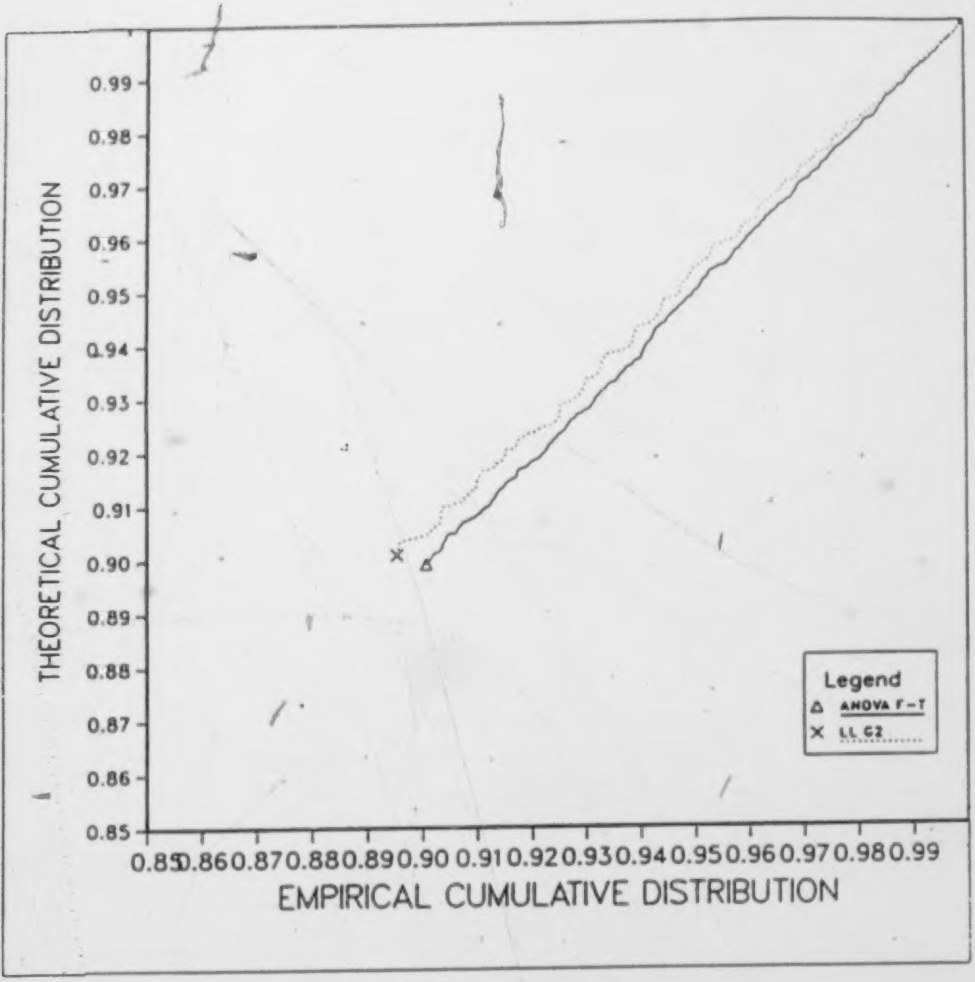

Figure A1.32 Conditional G^2 backward and ANOVA F with arcsine transformation testing for row effects: 2×2 table, 40 observations per cell, $p = 0.25$ (case 1.4).

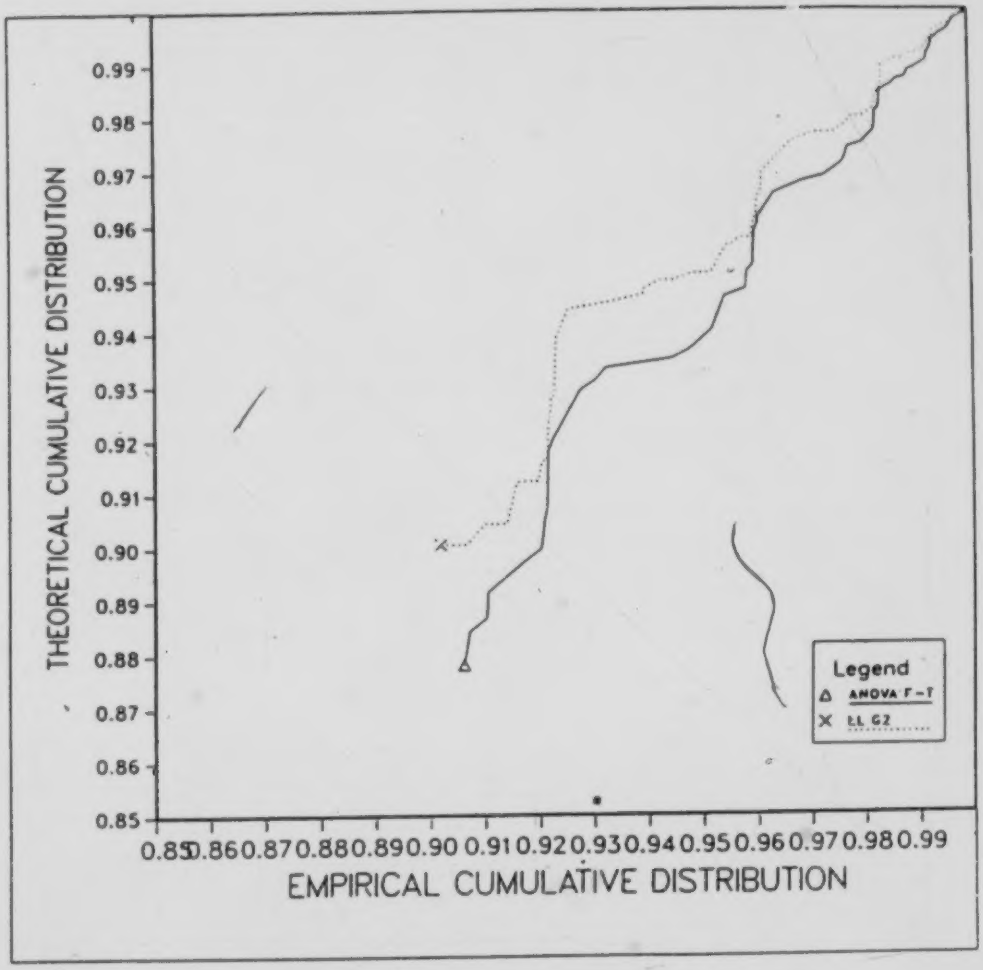

Figure A1.33 Conditional G^2 backward and ANOVA F for row effects: 2×2 table, 10 observations per cell, $p = 0.5$ (case 1.6).

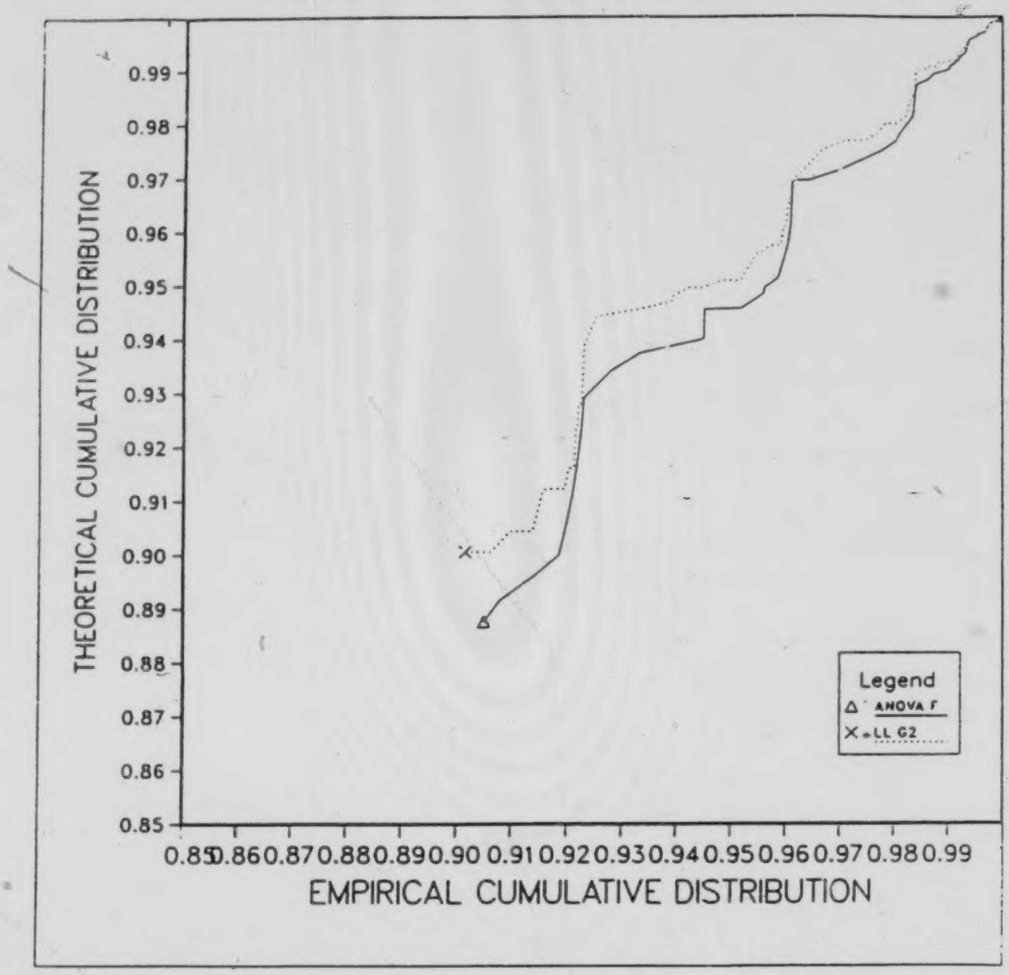

Figure A1.34 Conditional G^2 backward and ANOVA F with arcsine transformation testing for row effects: 2×2 table, 10 observations per cell, $p = 0.5$ (case 1.6).

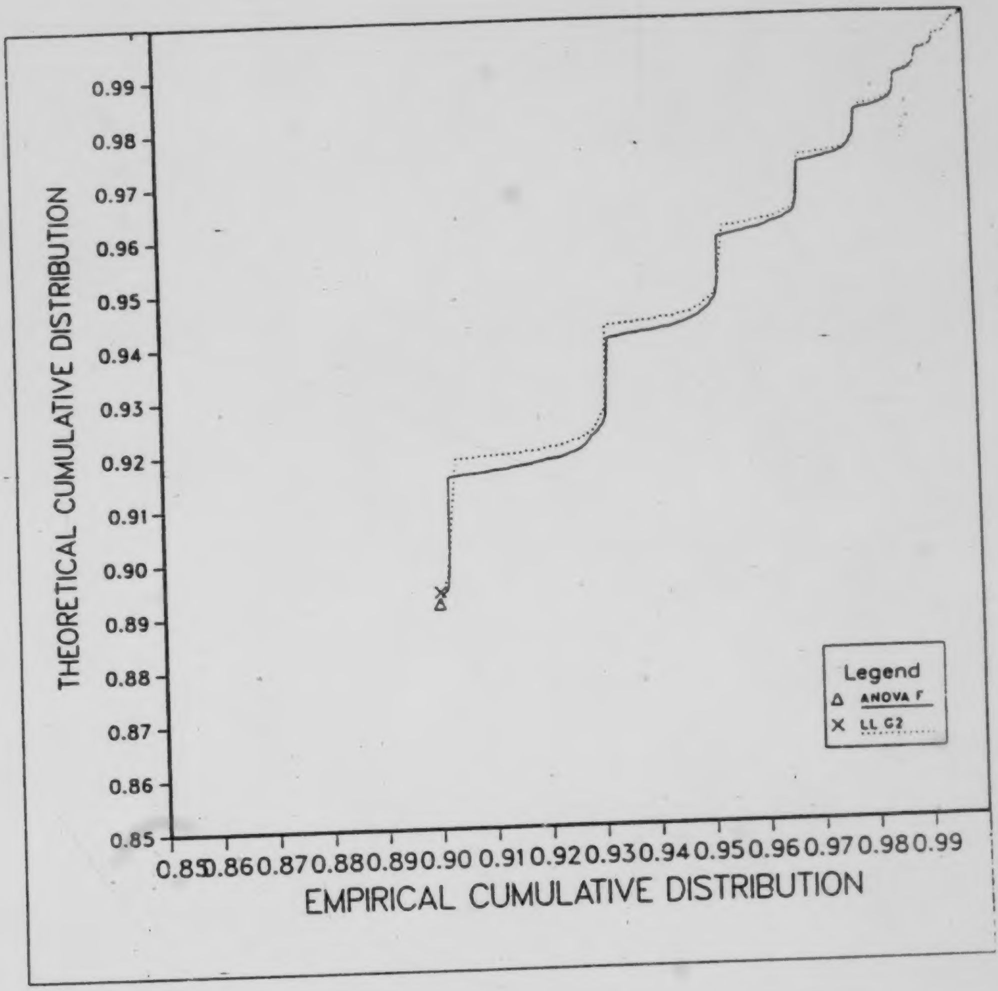

Figure A1.35 Conditional G^2 backward and ANOVA F for row effects: 2×2 table, 40 observations per cell, $p = 0.5$ (case 1.6).

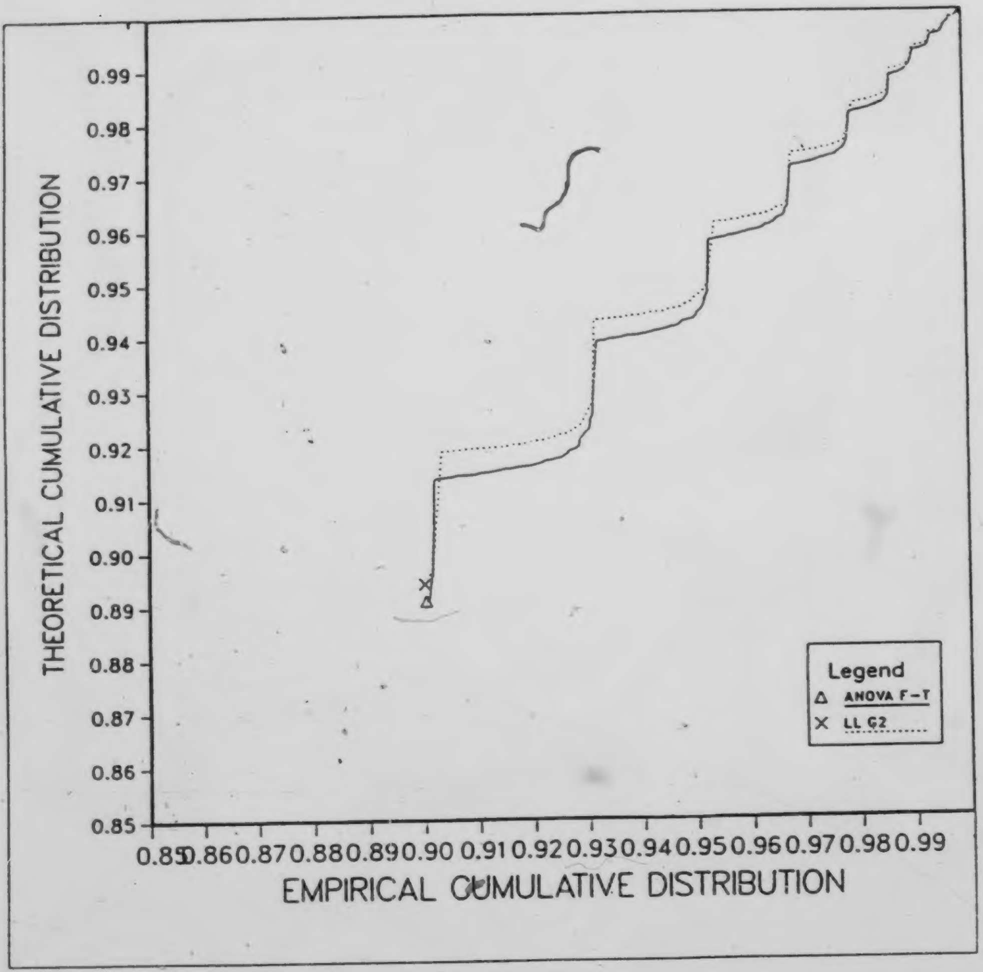

Figure A1.36 Conditional G^2 backward and ANOVA F with arcsine transformation testing for row effects: 2×2 table, 40 observations per cell, $p = 0.5$ (case 1.6).

Appendix A2

Additional Tables

EMPIRICAL TYPE I ERROR LEVELS

statistic	n per cell	nominal level (α)			
		0.10	0.05	0.025	0.01
F effect α	10	0.1058	0.0510	0.0205	0.0064
	10†	0.0865	0.0477	0.0237	0.0051
	40	0.0991	0.0493	0.0248	0.0098
	40†	0.1010	0.0494	0.0259	0.0112
F effect β	10	0.1036	0.0501	0.0214	0.0082
	10†	0.0815	0.0451	0.0224	0.0078
	40	0.1002	0.0472	0.0215	0.0091
	40†	0.1009	0.0466	0.0244	0.0119
F effect γ	10	0.1118	0.0514	0.0236	0.0082
	10†	0.0882	0.0481	0.0227	0.0082
	40	0.1012	0.0526	0.0269	0.0095
	40†	0.1038	0.0527	0.0271	0.0117
F effect αβ	10	0.1109	0.0544	0.0228	0.0077
	10†	0.0912	0.0505	0.0241	0.0074
	40	0.1022	0.0519	0.0256	0.0097
	40†	0.1008	0.0522	0.0277	0.0117
F effect αγ	10	0.1046	0.0497	0.0222	0.0077
	10†	0.0853	0.0467	0.0224	0.0070
	40	0.1022	0.0525	0.0240	0.0089
	40†	0.1012	0.0508	0.0247	0.0099
F effect βγ	10	0.1094	0.0540	0.0257	0.0096
	10†	0.0883	0.0502	0.0249	0.0082
	40	0.1043	0.0539	0.0264	0.0097
	40†	0.1036	0.0536	0.0273	0.0109
F effect αβγ	10	0.1060	0.0499	0.0213	0.0077
	10†	0.0867	0.0438	0.0215	0.0062
	40	0.1008	0.0510	0.0254	0.0102
	40†	0.1003	0.0532	0.0282	0.0114
LL NULL X^2	10	0.0869	0.0388	0.0230	0.0106
	10‡	0.0160	0.0045	0.0016	0.0005
	40	0.1000	0.0481	0.0247	0.0107
	40‡	0.0594	0.0263	0.0125	0.0050
LL NULL G^2	10	0.1439	0.0543	0.0237	0.0065
	10‡	0.0075	0.0022	0.0004	0.0000
	40	0.1359	0.0757	0.0427	0.0189
	40‡	0.0743	0.0328	0.0152	0.0046

† Arcsine transformation carried out before computation of statistic.

‡ 0.5 was added to each observation before computation of statistic.

Table A2.1 Type I error levels for case 2.1: Three factors, success probability of 0.1 per cell, 10 and 40 observations per cell.

observations per cell.

EMPIRICAL TYPE I ERROR LEVELS					
statistic	n per cell	nominal level (α)			
		0.10	**0.05**	**0.025**	**0.01**
F effect α	10	0.1011	0.0488	0.0237	0.0094
	10†	0.1058	0.0545	0.0268	0.0111
	40	0.1013	0.0483	0.0270	0.0107
	40†	0.0977	0.0472	0.0263	0.0101
F effect β	10	0.0986	0.0481	0.0236	0.0102
	10†	0.1037	0.0523	0.0266	0.0126
	40	0.1012	0.0502	0.0240	0.0105
	40†	0.0987	0.0480	0.0221	0.0094
F effect γ	10	0.1018	0.0535	0.0270	0.0100
	10†	0.1076	0.0583	0.0299	0.0131
	40	0.0974	0.0497	0.0250	0.0110
	40†	0.0958	0.0470	0.0243	0.0104
F effect $\alpha\beta$	10	0.0969	0.0457	0.0225	0.0099
	10†	0.1042	0.0519	0.0253	0.0112
	40	0.0986	0.0499	0.0241	0.0095
	40†	0.0983	0.0479	0.0222	0.0089
F effect $\alpha\gamma$	10	0.1027	0.0480	0.0237	0.0086
	10†	0.1052	0.0535	0.0256	0.0098
	40	0.1002	0.0504	0.0253	0.0102
	40†	0.0977	0.0497	0.0247	0.0094
F effect $\beta\gamma$	10	0.0973	0.0514	0.0277	0.0117
	10†	0.1058	0.0578	0.0320	0.0139
	40	0.1014	0.0500	0.0244	0.0093
	40†	0.0987	0.0488	0.0234	0.0089
F effect $\alpha\beta\gamma$	10	0.0964	0.0494	0.0245	0.0111
	10†	0.1028	0.0543	0.0281	0.0118
	40	0.1020	0.0520	0.0265	0.0094
	40†	0.0998	0.0508	0.0248	0.0087
LL NULL X^2	10	0.0980	0.0418	0.0193	0.0072
	10‡	0.0412	0.0161	0.0064	0.0013
	40	0.1001	0.0509	0.0269	0.0090
	40‡	0.0861	0.0425	0.0194	0.0072
LL NULL G^2	10	0.1613	0.0867	0.0429	0.0174
	10‡	0.0551	0.0209	0.0076	0.0022
	40	0.1099	0.0600	0.0323	0.0121
	40‡	0.0917	0.0492	0.0241	0.0074

† Arcsine transformation carried out before computation of statistic.

‡ 0.5 was added to each observation before computation of statistic.

Table A2.2 Type I error levels for case 2.2: three factors, success probability of 0.25 per cell, 10 and 40 observations per cell.

EMPIRICAL TYPE I ERROR LEVELS

statistic	n per cell	nominal level (α)			
		0.10	0.05	0.025	0.01
F effect α	10	0.0985	0.0570	0.0273	0.0128
	10†	0.0999	0.0501	0.0251	0.0119
	40	0.1076	0.0516	0.0244	0.0095
	40†	0.1028	0.0512	0.0237	0.0095
F effect β	10	0.0981	0.0587	0.0247	0.0109
	10†	0.0992	0.0501	0.0232	0.0106
	40	0.1027	0.0511	0.0251	0.0106
	40†	0.1077	0.0517	0.0224	0.0087
F effect γ	10	0.0964	0.0563	0.0258	0.0106
	10†	0.0964	0.0475	0.0230	0.0101
	40	0.1060	0.0498	0.0222	0.0091
	40†	0.1060	0.0498	0.0202	0.0081
F effect αβ	10	0.0991	0.0575	0.0269	0.0116
	10†	0.0997	0.0505	0.0244	0.0110
	40	0.1079	0.0525	0.0258	0.0121
	40†	0.1079	0.0526	0.0245	0.0107
F effect αγ	10	0.0937	0.0554	0.0254	0.0115
	10†	0.0946	0.0481	0.0232	0.0109
	40	0.1033	0.0507	0.0248	0.0096
	40†	0.1032	0.0508	0.0226	0.0083
F effect βγ	10	0.0917	0.0526	0.0222	0.0102
	10†	0.0932	0.0446	0.0203	0.0098
	40	0.1058	0.0514	0.0256	0.0115
	40†	0.1058	0.0515	0.0233	0.0098
F effect αβγ	10	0.0920	0.0534	0.0229	0.0096
	10†	0.0936	0.0455	0.0200	0.0089
	40	0.1049	0.0519	0.0244	0.0102
	40†	0.1050	0.0519	0.0226	0.0090
LL NULL X^2	10	0.1053	0.0494	0.0245	0.0087
	10‡	0.0676	0.0284	0.0129	0.0034
	40	0.1023	0.0509	0.0255	0.0096
	40‡	0.0952	0.0450	0.0221	0.0083
LL NULL G^2	10	0.1311	0.0754	0.0409	0.0190
	10‡	0.0891	0.0413	0.0208	0.0082
	40	0.1090	0.0564	0.0296	0.0117
	40‡	0.0997	0.0496	0.0249	0.0096

† Arcsine transformation carried out before computation of statistic.

‡ 0.5 was added to each observation before computation of statistic.

Table A2.3 Type I error levels for case 2.3: three factors, success probability of 0.5 per cell, 10 and 40 observations per cell.